



---

Statistical Policy  
Working Paper 19

**Computer Assisted  
Survey Information Collection**

---

Prepared by  
Subcommittee on Computer Assisted Survey Information Collection  
Federal Committee on Statistical Methodology

Statistical Policy Office  
Office of Information and Regulatory Affairs  
Office of Management and Budget

May 1990

MEMBERS OF THE FEDERAL COMMITTEE ON

STATISTICAL METHODOLOGY

(April 1990)

Maria E. Gonzalez (Chair)

Office of Management and Budget

Yvonne M. Bishop	Daniel Kasprzyk
Energy.Information Administration	Bureau of the Census
	Daniel Melnick
Warren L. Buckler	National Science Foundation
Social Security Administration	
	Robert P. Parker
Charles E. Caudill	Bureau of Economic Analysis
National Agricultural Statistical Service	David A. Pierce
	Federal Reserve Board
John E. Cremeans	
Office of Business Analysis	Thomas J. Plewes
	Bureau of Labor Statistics
Zahava D. Doering	
Smithsonian Institution	Wesley L. Schaible
	Bureau of Labor Statistics
Joseph K. Garrett	

Bureau of the Census

Fritz J. Scheuren

Internal Revenue Service

Robert M. Groves

Bureau of the Census

Monroe G. Sirken

National Center for Health

C. Terry Ireland

Statistics

National Computer Security

Center

Robert D. Tortora

Bureau of the Census

Charles D. Jones

Bureau of the Census

#### PREFACE

The Federal Committee on Statistical Methodology was organized by the

office of Management and Budget (OMB) in 1975 to investigate methodological issues in Federal statistics. Members of the committee, selected by OMB on the basis of their individual expertise and interest in statistical methods, serve in their personal capacity rather than as agency representatives. The committee conducts its work through subcommittees that are organized to study particular issues and that are open to any Federal employee who wishes to participate in the studies. Statistical Policy Working Papers are prepared by the subcommittee members and reflect only their individual and collective ideas.

The Subcommittee on Computer Assisted Survey information Collection investigated the use of computers in collecting survey information. This report covers the different ways in which small computers can be used to improve data collection.- For example, the report describes computer assisted telephone interviewing (CATI), computer assisted personal interviewing (CAPI), data collection using touchtone telephones, and voice recognition. More than most working papers the relevance of the information in this report will age very quickly.

Various methodological issues are also addressed in this report. For example, issues discussed include human-machine interfaces, software development, hardware planning, and computer security.

The Subcommittee on Computer Assisted Survey Information Collection was chaired by Terry Ireland of the National Computer Security Center, Department of Defense.

i

#### CASIC Subcommittee Members

C. Terrence Ireland, Chair

National Computer Security Center

(Defense)

Thomas Anastasio

National Computer Security Center

(Defense)

Martin Baum

National Center for Health Statistics

(Health and Human Services)

William Blackmore

Energy Information Administration

(Energy)

Richard Clayton

Bureau of Labor Statistics (Labor)

Ann Ducca

Energy Information Administration

(Energy)

Ralph Gillman

Energy Information Administration

(Energy)

Maria E. Gonzalez, Ex officio

Office of Management and Budget

(Executive Office of the President)

Stuart Katzke

National Institute of Standards and Technology

(Commerce)

George Kraft

National Institute of Standards and Technology

(Commerce)

Cathy Mazur

National Agricultural Statistical Service

(Agriculture)

John Sietsema

National Center for Education Statistics

(Education)

## Acknowledgments

The idea to develop a Statistical Working Paper on the use of computers to support the collection of survey information was first put forward by Yvonne Bishop of the Energy Information Administration. Ms. Bishop has a special interest in data collection techniques that do not involve an interviewer. With the advice of members of the Federal Committee on Statistical Methodology (FCSM), Maria Gonzalez organized a subcommittee with an expanded scope to examine a range of computer methodologies that supported the collection of information, the subcommittee on Computer Assisted Survey Information collection, (casic) . The members of the CASIC Subcommittee further expanded the report to include the three important methods of data collection, Computer Assisted Telephone interviewing (CATI), Computer Assisted Personal Interviewing (CAPI), and Computer Assisted Self Interviewing (CASI). For each related technological area from software interfaces to computer security, the, CASIC Subcommittee investigated and wrote



sections of the working paper that showed the application of these areas to CATI, CAPI, and CASI. The CASIC Subcommittee thanks the members of the FCSM for their advice and comments on several drafts of the working paper. Special thanks go to Charles Caudell (HASS) and Joe Garrett (Census). for their in depth comments on the various drafts.

iii

COMPUTER ASSISTED SURVEY INFORMATION COLLECTION (CASIC)

TABLE OF CONTENTS

Part I. Executive Summary	1
A. Introduction	1
B. Computer Assisted Survey Information Collection	2
Part II. Introduction	3
A. Objectives, Scope, and Users	3
B. Federal Information Processing Standards	8

C. Organization of Report	9
Part III. Options for Automated Statistical Surveys	11
A. Computer Assisted Telephone Interviewing (CATI)	11
B. Computer Assisted Personal Interviewing (CAPI)	15
C. Computer Assisted Self Interviewing (CASI)	17
Part IV. Methodological Issues	25
A. Human-machine Interfaces	25
B. Software Development	32
C. Data Collection Programs	36
D. System Interfaces For Data Conversion	41
E. Computer Security	44
F. Hardware Planning	50
G. Network Planning	54
Part V. References	63
Part VI. Appendices	67
A. Costs	67

B. Quality Improvements offered by CASIC	73
C. Survey Examples	78
D. Taxonomy	94
E. Glossary	96

v

## I. Executive Summary

### I.A. Introduction

Surveys have used computers since the Bureau of the Census obtained the UNIVAC I. Since that breakthrough,, the power of rapid calculating has been applied to almost every phase of the survey process, including sample design, sample selection, and estimation. The most important implication of these applications is that survey practitioners can now consider a growing range of techniques that were not affordable, or even thought of, before the availability of inexpensive and fast calculating capability.

The last major survey operation to benefit from automation is data collection. Computers were first applied to collection using mainframes to control certain aspects of telephone collection, and computer Assisted Telephone Interviewing (CATI) was born. The first applications of CATI provided a flood of research worldwide evaluating the impact of this technique on the survey error profile and costs. CATI is now used to help interviewers in all collection activities, including scheduling calls, controlling detailed interview branching, editing and reconciliation, thus providing much greater control over the collection process and reducing many sources of error. Simultaneously, a tremendous storehouse of information is captured by the computer to provide additional insight into the data collection process.

In just two decades, CATI has become a standard collection vehicle grounded strongly in a firm body of research.

The ongoing advances in computer technology, and particularly the arrival of microcomputers, continue to offer survey practitioners more fertile ground for improving the quality of published data. The first

portable computers were quickly pressed into service to duplicate the advantages of CATI in a personal visit environment. Thus, Computer Assisted Personal Interviewing (CAPI) grew from the seeds of CATI.

While CATI and CAPI represent advances for surveys requiring interviewers, microcomputers are now finding important roles in self-administered questionnaires, where interviewers are not needed. These roles take advantage of more advanced technology and the widespread availability of technology to allow respondents to complete the questionnaire without the assistance of an interviewer. Prepared Data Entry (PDE) allows respondents that have a compatible microcomputer or terminal to access and complete the questionnaire directly on their screen.

Touchtone Data Entry (TDE) allows respondents to call and answer questions posed by a computer using the keypad of their touchtone telephone for well-controlled and inexpensive collection. As an extension of this approach, recently developed techniques in

Voice Recognition Entry (VRE) allow respondents to answer questions by speaking directly into the telephone. The computer translates the respondent's answers into text for verification with the respondent and then stores the text in a data base.

These and other collection methods will continue to evolve out of the work now underway. New technology will assuredly bring more options for survey practitioners to consider.

The use of these collection methods, while bringing needed improvements in the quality of collected data, has created other challenges. These automated collection methods are made possible through the close interaction of statisticians, subject matter experts and colleagues in the computer sciences. To use these methods effectively, each profession must learn and use the models and techniques of the other professions. This close relationship will

continue to grow, with advances in each field supporting advances in the others.

The goal of this report is to profile several automated survey collection methodologies and provide a glimpse of what future technological advances may offer to survey operations.

The selection of one or more of these collection methods depends on a clear understanding of computer applications. Software and hardware selection can be essential to success,, as may be the use of networks for the computers. As with any survey method, the need to assure the confidentiality of the data gathered and stored by the computers is critical.

This report discusses several data collection methodologies now being used in Federal agencies in terms of procedures, impact on quality and costs. It also discusses the significant issues surrounding the use of advanced technologies to augment survey data collection.

## I.B. Computer Assisted Survey Information Collection (CASIC)

For this report, the Subcommittee defines Computer Assisted Survey information Collection as those information gathering activities using computers as a major feature in the collection of data from respondents, and in transmitting of data to other sites for post-collection processing. It is in this area of survey operations that technology is now having the greatest impact.

## II. Introduction

### II.A. Objectives, Scope, and Users

The Subcommittee on Computer Assisted Survey Information Collection was established in October 1988 to document and discuss the



status and potential use of advanced technology for collecting statistical data, for its transmittal to central processing sites, and the conceptual and practical issues surrounding implementation. High quality published data begins with collecting high quality data from respondents. Much of survey processing addresses, and compensates for, weaknesses in the quality of the collected data and absence of uncollected data. The survey questionnaire, received on time, completely filled out and accurate, can reduce post-collection errors and their related costs.

The Computer Assisted Survey Information Collection Subcommittee of the Federal Committee on Statistical Methodology has studied the various implications of the vast computing power now available to support statistical surveys and is providing this information for use throughout the Federal Government.

#### Objectives

The primary objective is to describe emerging methods of interactive electronic data collection and transmission, potential

benefits, and current examples of their use in Federal surveys. This report also covers techniques and appropriate references to the literature.

A secondary objective is to consider specific methodologies and related issues stemming from the use of computer assisted statistical surveys. Also addressed are other practical considerations involving human-machine interfaces, software design, hardware features, data transmission and computer security. The issues involve such factors as quality, costs, and respondent reaction to computerized surveys.

Some advantages of automated surveys are:

- a. improved data quality from (1) the introduction of automated questionnaire branching, editing features, and computer utility support; and (2) a shorter processing path from data collection to data processing (e.g., reduced keying errors because keying of the paper questionnaire is no longer necessary).

b. improved timeliness of data capture by the elimination. of  
some data entry steps and of extensive editing.

3

c. increased flexibility in data gathering (e.g., for  
conducting multiple version questionnaire surveys involving  
question reordering and different natural languages).

In deciding which collection method to use, quality is a relative  
idea that is affected by a tradeoff between cost and benefit. The  
choice of a data collection method is usually based on a combination  
of performance and cost factors. Together they determine affordable  
quality. For traditional collection methods, these factors and the  
decision-making process are usually well-known. Now, as technology  
progresses, new methods are being tested that expand the array of  
potential collection tools and challenge the survey-designer to

reevaluate old cost/performance assumptions.

These semi-automated collection applications fall naturally into 3 areas: (1) Computer Assisted Telephone Interviewing (CATI) where the interviewer and respondent talk over a telephone, limiting their personal interactions while maintaining the substantial flexibility provided by a telephone; (2) Computer Assisted Personal Interviewing (CAPI) where the interviewer and respondent talk directly across the table, although this direct access comes with the cost of additional logistical problems; and (3) Computer Assisted Self Interviewing (CASI), a newly coined phrase to describe situations where the interviewer is replaced by interaction with the computer. -

Subcategories include Prepared Data Entry (PDE) where the respondent uses a computer terminal; and Touchtone Data Entry (TDE) and more recently, Voice Recognition Entry (VRE) where the respondent interacts with a computer over a phone line.

However, computer applications are not limited to obtaining data from respondents. In addition, the prompt transmittal of reported data to the processing facility and the conversion of data

to proper formats are important to the publication of timely and relevant information.

New options will encourage reconsideration of old assumptions about quality, cost, and technology. Decisions made years ago in an era of fewer alternatives should be reviewed periodically. Many factors can change in a short period. Only a few years ago, automation costs were driven by the scarcity of mainframe hardware capacity. Now the labor involved in developing specialized systems dominates automation costs. Portable and desktop microcomputers were not widely available at the beginning of this decade. Now, widely available, inexpensive and powerful, they are an assumed part of the work environment. The tough questions involve the selection of the appropriate system configuration.

The general goal of this report is to challenge Federal survey managers to reconsider their operations in light of recent changes

in survey methods available, or attainable through new technology, and to reassess their methods of providing information to the public that is accurate, timely and relevant.

#### Scope

Automated data collection includes three major groups of people: the respondents, the interviewers, and the designers and developers of the system and procedures for collection. This report covers the essential factors involved in successfully including the requirements of each group.

The survey operations considered in this report include the computer-related activities of design - and development of the

questionnaire, interviewing, data entry, editing and follow-up for nonresponse or edit reconciliation, data transmission and data conversion.

The critical activities of sample design, sample selection and estimation are not included in the scope of this report. Still, the choice of an automated collection method is important to these activities. This choice must be an integral part of the survey design. For example, the decision to use CATI to improve collection of time critical data may provide the sample designer with additional flexibility to consider techniques that require rigorous sample control or complex questionnaire branching logic.

## Respondents

The respondent must be considered the primary user of any survey vehicle, whether automated or not, and all aspects of the response environment must be developed with the respondent in mind.

The cooperation of respondents is the single most critical factor in survey operations, and they must be treated with the greatest care. Even one-time surveys must strive to leave the respondent with the

feeling of contribution and importance, and the willingness to participate in future surveys. Thus, our primary job is to consider computer-related techniques that allow the respondent to answer the survey completely and accurately in a natural environment.

Automated collection methods provide survey managers with opportunities to improve control and reduce sources of error. These methods also can be designed to capture workload and performance data in the background while interviews are conducted. However, these - features must not interfere with the natural interactions during the interview.

The transition to automated surveys presents additional challenges. For example, in a switch from mail questionnaires to CATI, the surveyor must work with the respondents to remove their



uncertainties about the transition in order to retain their continuing cooperation.

The arrival of a variety of automated self-response methods involving computerized questionnaires presents new challenges for ensuring that the respondent is sufficiently knowledgeable and comfortable dealing directly with the computer. As always, the respondent must be trained in the use of the collection process. Whether by simple instructions or more formal procedures manuals, the surveyor must work diligently to develop simple, clear directions for use, or risk losing the full cooperation of the respondent. For example, in the use of PDE, respondents must interact directly with computer displays. This requires understandable questions, adequate help facilities, and a clear set of allowable answers. Finally, just as managers must worry about interviewers' illness, absence, vacations, and vacancies, designers of automated self-response systems must include emergency back-up procedure to assure that respondents can complete the survey.

The design of the human-machine interface requires a clear

understanding of what the respondent expects. Do people react to questions differently when presented on paper compared to telephone interviewers and still differently if posed from computerized displays or computerized voices? Also, what information is lost by changing from personal visits, where the interviewer can assess a variety of non-verbal clues, to telephone collection, or automated self-response where voices are not directly heard? What are the differences in application of these techniques in household versus establishment surveys?

While new automated methods provide many features attractive to survey designers, new responsibilities come with their use. The respondent must be assured of the confidentiality of the data provided. Confidentiality is the cornerstone of respondent cooperation, from the interview through final processing, estimation, and storage of microdata. Whereas face-to-face interviews provide an environment where the respondent can assess and control access by others, use of telephone collection and transmission of self-reported data creates new problems in confidentiality. The integrity and authenticity of the respondents answers during the transmission

process is a related issue. The ability to transmit large volumes of data from remote sites may only partially solve collection, problems in some surveys that require actual signatures and protection of the transmitted data.

#### Interviewer

The second most important user is the interviewer. The systems provided to help in the interview process must be easy to use, must work consistently and must provide improvements in the interview environment. Early use of CAPI required interviewers to

carry the first generation of portable computers to the respondent's home. These heavy machines were often left in automobiles until the interviewer could decide that the respondent was home. The result was

reduced productivity and higher costs.

Interviewers must believe that computer assistance will improve their effectiveness. They need to be convinced that the computer is simply a tool to speed and simplify their work. CATI, CAPI and CASI support specific wording for each question, and simplify moving to the next question, which is often dependent on previous answers. However, these systems can be over-developed so that interviewers are left little or no discretion for judgment or contribution. The result may be low morale, indifference, deviation from established procedures, and high turnover rates.

#### System Designers

The third important user is the system designer who may use the computer environment to design the survey and to lay out the procedures for its use. Besides the ease of use to both respondent and interviewer, the decisions made early in the development process carry over to the ongoing use and maintenance of the system for years. The design environment is similar to that used in any software

development process. Software tools that support this "software engineering" process should give flexibility to the designer and provide, for long-term maintenance of the survey.

System designers face difficult choices, such as building customized systems from scratch versus linking standardized "off the shelf" software packages. The inevitable limitations must be compared against reduced maintenance and lower start up costs.

## II.B. Federal Information Processing Standards

Today, more than ever, information is the force that drives the activities of the Federal Government and information processing systems are the mechanisms that process, store, and transfer this information. Information processing standards play an increasingly

important role in the strategies of Federal Agencies to make more effective use of their information processing systems by providing needed interoperability of systems and equipment, portability of data and software, and methods for protecting data and computers from accidental and intentional harmful events. CASIC systems, like other Federal information processing systems will be more effective if they implement standards that provide for interoperability, portability, and security.

Within the Federal Government, the National Institute of Standards and Technology (NIST) has the responsibility of promulgating Federal Information Processing Standards and Guidelines for hardware, software engineering, electronic document interchange, data management, ADP operations, computer security, and ADP related telecommunications. In addition, NIST develops conformance tests for its standards where appropriate. Developers of computer assisted statistical survey systems should use NIST's standards and guidelines whenever possible during - the design, implementation, and operation of their systems. A reference to NIST's standards program and available standards and guidelines can be found in Section V under the

heading of "Standards." Additional information about NIST's program

may be obtained from:

Program Coordination and Support Group

National Computer Systems Laboratory

Building 225, Room B151

National Institute of Standards and Technology

Gaithersburg, MD 20899

Telephone: (301) 975-2833

## II.C. organization of the Report

This report is intended to provide reference and guidance for survey practitioners across the Federal Government in planning and refining data collection methods. By sharing information and experiences, others may gain and add to the effectiveness of

governmental survey activities. The potential audience is much broader than those involved in statistical surveys. Many of the methods described and the technological issues discussed are applicable to any information collection activity, including the collection of management information, program cost, productivity,, and workload data.

Part III covers the 3 major areas of CATI, CATI, and CASI where the computer supports survey information collection. Each major application is defined and current survey application experiences are described. Each discussion describes the impact on specific survey error components and potential for future applications.

Part IV provides a discussion of broad technological and developmental issues in the use of computer assisted surveys. The areas selected for consideration are: the human-machine interface; software development; data collection systems; systems interfaces for data conversion; computer security; hardware planning; and network planning which includes electronic mail.



Part V contains references organized by categories consistent with the organization of the report.

Part VI contains the appendices. Appendix VI.A provides a discussion of cost measurement relating to use of computers to collect survey information. Appendix VI.B provides a general discussion of the improvements of quality that can be expected with the use of computers. Appendix VI.C provides a series of survey efforts currently underway, with a point of contact for additional information. Appendix VI.D lays out a suggested classification model for surveys that depend on computer support. It is consistent with the various models in the body of this report. Appendix VI.E contains a glossary of words in active use where computers and surveys come together.

### III.A. Computer Assisted Telephone Interviewing (CATI)

#### Definition

Computer Assisted Telephone Interviewing or CATI is a computer assisted survey process that uses the telephone for voice communications between the interviewer and the respondent.

CATI replaces the traditional paper-and-pencil questionnaire interviewing. The questionnaire is displayed to the interviewer by the computer who then relays the question over the telephone to the respondent. The answers are given to the interviewer for entry into the computer. The collections of questions are structured so that computer examination of previous answers can be used to select the next question in sequence. Computer-generated help facilities can be initiated by the interviewer on command.

The interview environment can be computer generated or handled manually by the interviewer. As the CATI systems grow in

sophistication, many manual functions will be taken over by the computer: sampling unit selection, scheduling of telephone calls, automatic dialing, and callbacks to respondents who are not reached on the initial call.

Data collected by CATI should have significantly fewer errors than manual methods because the interviewer can validate directly respondent's data that fails internal and historical edit checks. Time and cost requirements for data collection, validation, and data conversion should be reduced. Computer controlled questionnaires make it possible to use more sophisticated designs than can be administered with paper-and-pencil forms. They can include complex logic structures and questions finely tailored to the circumstances associated with a specific sampling unit.

#### Examples of Current Use

The exact number of CATI installations throughout the world is unknown. It probably is more than 1,000 considering the number of countries, universities, and private sector vendors and survey

research installations involved in surveys. In 1988, the U.S.

Government had 51 cooperating CATI centers.

Both opinion and factual data are collected using CATI. Most questionnaires contain a mix of these data types. Questionnaires range from several questions with very little data validation to several hundred questions customized for specific respondents providing the ability to collect conveniently the same data in different respondent environments.

The National Agricultural Statistics Service (KASS) within the United States Department of Agriculture (USDA) executed its first CATI questionnaire (Multiple Frame Cattle Survey) during 1982 in California using four workstations and completing 100 interviews. The

questionnaire consisted of 41 questions. Today the largest known CATI questionnaire is the December Agricultural Survey. it is used in 14 states with questionnaires customized for each state. This survey has over 200 questions with production items recorded in units convenient to the respondent and converted to a common unit for data validation and recording purposes.

Today, HASS conducts a total of nine recurring CATI surveys. The surveys are monthly, quarterly and annual. In 1988, NASS completed 125,000 CATI interviews using 183 data collection work stations in 14 remote sites located in state statistical offices. Besides the recurring CATI activity, NASS conducted three special data collections in 1988 and two already were scheduled for 1989. The questionnaires were developed over a very short period. Training time was short. The data collection period was somewhat short (3 days - 2 weeks). NASS found that CATI lends itself very well to applications with short implementation schedules. Field testing of the questionnaires is efficient because once a problem area is identified, the questionnaire can be modified and tested on another respondent in generally less than an hour.

Also, the Bureau of Labor Statistics (BLS) currently uses CATI in 17 States to collect monthly data on employment, hours and earnings from 6,000 respondents. BLS further uses CATI (1) to collect Consumer Price Index (CPI) housing data; (2) to collect hours at work and hours paid as an input to productivity measures; and (3) for special purpose studies to support Department of Labor initiatives. In addition, BLS uses CATI methods to conduct telephone record check surveys to improve data quality.

#### Computing Environment

The Uses of CATI are limited only by the capability of telephone technology and the use of personal interviewers. CATI is one of several phases of the total data collection process. It can be used for nonresponse follow-up where initial contact is made by CATI, mail or capi.

The ability to use varied data collection techniques is contingent upon the ability to develop computer questionnaires with

common software that can support the various data collection options.

Common software is important to assure the same data is collected and the same validations are applied.

The computer has to be responsive in delivering sample units and questions-to the interviewer. The computer response times for both interviewer and respondent must be less than what they would

12

perceive as an unnecessary delay. For example, experience has shown that longer than a second between questions is too long for an impatient respondent. Longer than half a second wait for the display of the next question is too long for the interviewer. During this period the computer may be required to access several databases and do complex mathematical computations which would include logical decisions affecting subsequent questions.

The computer must deliver a different sampling unit in less than 10 seconds, and ideally in less than five. During this period the machine may have to query several potential respondent queues that relate to scheduled callbacks in different time zones,- to previous busy signals to be retried every 15 minutes; to special handling of specific respondents by specific interviewers; to the generation of new sampling units; and to the disposition of the completed interview as correct.

The software that drives the questionnaire must be easy for the interviewer to use. Question paths through a questionnaire must be simple and easy for the interviewer to handle. Menus with abbreviated questions or questionnaire areas are desirable. Skipping back to an earlier question, changing that answer, and establishing another route through the questionnaire must be easy and quick to do. Commands must be standardized for use in related surveys to enable "second nature" reactions by the interviewer in any given situation.

The design of a CATI questionnaire poses problems beyond the design of standard questionnaires. If the designer has problems



developing the questionnaire, the interviewers will almost surely find it difficult to use. The objectives of the survey questions in a computerized questionnaire may be no more complex than questions used in pencil-and-paper surveys. However, the flexibility provided by automated question paths makes their design more difficult as the possible sequences of questions must be worked out during design. Paths and branching must be worked out in advance and there may be significant differences in question wording and in their number. Automatic sampling unit management can pose some difficult logic problems for the automated survey designer. Data validation using historical or internal data correlations is a complex logic problem, but is essential for recurring surveys. Well designed computer environments provide the interviewer with the ability to review the respondent's answers for correctness and to annotate unusual circumstances.

Before the computer questionnaire designer can begin, the questions must be developed by the survey staff using knowledge of statistical theory and specific subject matter. This survey staff also must be well versed in face-to-face, self-administered, and

telephone questionnaire design. In the face-to-face interview the interviewer can offer explanations of the question, then probe for additional information; and if necessary,, provide the respondent

13

with the paper version of the questionnaire. The respondent can study, read ahead, reflect, and finally answer with a clear understanding of the meaning of the question.

For a self-administered questionnaire, the respondent no longer has the benefit of the interviewer, but still can examine the questionnaire in detail. in telephone interviewing the respondent may not have the form in-hand and thus may be missing the visual clues needed to understand the question. Therefore, questions used in telephone interviewing should be structured using single concept questions. Some simple applications rely less on posing very

structured questions and more on a "forms-screen" approach. This approach replicates the survey form on the computer screen. Edit failures may be highlighted, perhaps with a different color, and the interviewer is trained to ask probing questions to reconcile suspected inconsistencies in the responses.

### III.B. Computer Assisted Personal Interviewing (CAPI)

#### Definition

Computer Assisted Personal Interviewing (CAPI) is a personal interview conducted usually at the home or business of the respondent using a portable personal computer. In many respects it differs from CATI only in the presence in the same room of the interviewer and the respondent. As with CATI, the questionnaire is programmed into the

computer with all the necessary logic to control the question path -- the logical flow of the questions based on such factors as previous answers -- and provides both for computer generated editing by pointing out inconsistencies to the interviewer and for direct editing by the interviewer. The system must be self-contained as the interviewer does not have immediate access to supervisory assistance or to other data sources. The interviewer reads aloud each question as it appears on the screen and records the respondent's answer in the computer while providing interactive assistance to the respondent.

#### Examples of Current Use

CAPI is currently being used by the National Center for Health Statistics (NCHS) for the implementation of the National Health Interview Survey (NHIS). The Census Bureau is performing the field data collection for NCHS. The NHIS is a household survey conducted in approximately 50,000 households per year. CAPI has been used to collect a portion of the survey data: the AIDS supplement questionnaire that requires approximately 15 minutes to complete. The 1990 Health Promotion and Disease Prevention Questionnaire of the NHIS

will be fielded in January 1990. Major tests of CAPI have been conducted by the Bureau of the Census and the Research Triangle Institute. National Analysts conducted a nationwide CAPI for the USDA sponsored 1987 Nationwide Food Consumption Survey. The Bureau of Labor Statistics used CAPI for establishment record check surveys. National opinion Research Center also is experimenting with CAPI. In Europe, CAPI has been used by the Netherlands Central Bureau of Statistics to collect data for the Netherlands Labor Force Survey. The U.K. Office of Population Censuses and Surveys has also carried out a major test of CAPI. Most of these efforts are at an early stage of CAPI development.

#### Potential Uses

CAPI can be used for all household surveys and establishment surveys, and the software can be used for any of the other automated data collection mechanisms. As the technology improves to provide lighter computers with longer battery life and user friendly software, CAPI will be used more often, particularly for quick turnaround surveys. Procedures for developing CAPI

questionnaires are similar to those for CATI. However, greater emphasis must be placed on help features because the CAPI interviewer cannot rely on nearby experts.

The type of resources and expertise needed to apply CAPI technology to a survey are dependent on the availability of a good authoring system. If an authoring system is readily available, the CAPI, survey instrument can be prepared by the typical survey instrument designer with little or no computer experience. Computer programming assistance will be needed to write the case management and output portions of the software. Usually these portions of the survey vary with each survey or survey instrument; therefore they must be custom programmed. On the other hand, if an authoring system is not available, the entire CAPI instrument must be custom programmed with

either a general purpose language or a special purpose CAPI language.

In either case, computer programming expertise is required. The level of expertise is dependent on the language selected. in addition, the survey instrument preparation will require the services of a survey instrument designer who will need to work very closely with the computer programmers.

### III.C. Computer Assisted Self Interviewing (CASI)

#### Definition

Computer Assisted Self Interviewing (CASI) has been introduced into this report as a category to cover a new but growing area of computer assisted surveys that involves data collection without the direct presence of an interviewer. CASI can take several different

forms that are differentiated by the collection method. These include Prepared Data Entry (PDE) where the respondent answers questions displayed on a computer terminal; Touchtone Data Entry (TDE) where the respondent answers computer generated questions by pressing buttons on a telephone; and Voice Recognition Entry (VRE) where the respondent answers questions by speaking directly into a telephone. We consider each in turn.

## Background

Self-response data collection has always been used for many surveys that are mailed out. This form of self-response collection features simplicity in administration leading to low initial overhead when compared to CATI and CAPI. However, mail self-response necessarily involves a reduction in control over the collection process. It is difficult for the survey practitioner to assess the status of the collection effort, e.g., whether the responses are in transit or still in the respondents' hands. Extensive mail or telephone follow-up involves great costs, perhaps offsetting the



original simplicity of mail, and-risks ongoing cooperation, especially if the response is "in the mail."

In annual or quarterly surveys, mail may be the appropriate vehicle. In time critical surveys, the characteristics of mail collection leave wide gaps in control. Computer Assisted Self-Response methods now being introduced into surveys hold great promise to maintain the advantages of mail self-response, while improving control and the ability to intervene in the collection process.

Definition - Prepared Data Entry (PDE)

Prepared Data Entry (PDE) places the respondent in direct contact with a computerized questionnaire through a computer terminal. in a sense the computer is acting as the interviewer in a manner similar to CATI or CAPI interviewers.

The respondent uses a personal computer or terminal to fill out interactively the survey questionnaire. As each item appears on

screen, instructions and definitions for that item appear on a split screen or are accessible by pressing a help key. As data are entered, range and consistency checks are automatically applied and

17

anomalies pointed out to the respondent. The response to previous items may control the question path of the questionnaire. Because of the lack of an interviewer to help the respondent, the guidance provided by the program must be substantial and the computer literacy of the respondent is essential, at least at this stage of development.

This category of automated data collection programs includes a rapidly expanding set of respondent initiated data entry and transmission methods. These methods are directly dependent upon the computer and telecommunications hardware available to the data providers. Individuals, small businesses, or reporting agents can enter data into a personal computer in response to pre-programmed

floppy disks and mail the disks to the collecting agency. Firms with modems can transmit the data through telephone lines directly to the collecting agency's mainframe, or via an electronic mail service. Larger firms with mainframes can download the data to a PC, then either transmit directly from the PC over a modem to the agency's mainframe or place the data on a diskette and mail it to the agency.

These methods eliminate the need for rekeying the data and suffering the risk of data entry errors. The transmission methods using telephone lines save several days in each collection cycle by eliminating dependence on the physical transportation of machine-readable data whether by mail or special couriers. The data must be checked to detect and correct errors introduced during transmission.

#### Examples of Current Use

In the early 1980's, the Internal Revenue Service (IRS) decided that the electronic transmission of returns by tax preparers to IRS would be both a practical and cost-beneficial alternative to the mailing of paper tax returns when a refund is claimed. According to the Agency, the benefits of electronic filing would include: (1)

reduced manual labor costs required to process, store, and retrieve returns; (2) faster processing and retrieval of tax data; and (3) reduced interest IRS must pay to taxpayers who file timely refund returns that are not processed on time by the IRS.

Further, IRS reports show that electronically transmitted returns are processed with significantly fewer errors than paper returns. According to IRS figures for the 1988 filing season, as of April 29, 1988, 20 percent of paper returns processed by IRS had errors and only 5.5 percent of those filed electronically had errors. For taxpayers, electronic filing can mean refunds up to 3 weeks sooner, and because IRS can deposit these refunds directly into taxpayer bank accounts, refunds may arrive 3 to 4 days earlier

than that. For tax preparers, the ability to provide electronic filing services to taxpayers promises a competitive business edge.

The Petroleum Supply Division (PSD) of the Energy Information Administration (EIA) decided in 1987 to investigate electronic forms submission to collect the Petroleum Supply Reporting System (PSRS) survey forms. Ten of the major petroleum companies who file the mandatory "Monthly Refinery Report" were contacted to assess their PC and communications capabilities. The respondents contacted showed interest in investigating the use of PC's to collect this data. Often they were already using PC's for business, personal or academic purposes. The respondents either had a PC in their office area or had access to one in another office. Software such as Lotus 1-2-3 and dbaseIII could usually be found on these PC'S. Some PC's were equipped with communications capabilities and those respondents were already using telephone lines for company reporting. It appeared to be the appropriate time for the PC to enter the PSRS data collection process.

Early in 1988, PSD developed the Petroleum Electronic Data

Reporting option (PEDRO) and began providing its respondents with a software diskette by which they could create an electronic image of the form on a PC screen and enter their data in the appropriate cells. Firms having the necessary software capabilities can use their data base to feed data directly to the electronic survey form eliminating keying and transcription errors. User-friendly software with help functions has been added to data entry functions to provide quick reference to definitions, conversion factors or other information to speed the completion of the survey form. This eliminates the need to search hard-copy files for survey forms instructions, product definitions, conversion tables, etc.

Definition -- Touchtone Data Entry

Touchtone Data Entry (TDE) has been used for many years in the private sector for a growing range of applications. TDE, also known as voice response, is used for banking by telephone, call routing, college class registration and "talking yellow pages" to name just a few. The process is simple. The caller initiates a call to a computer which asks a series of questions. The caller answers using

the touchtone keypad and the tones are recognized by the computer.

The process offers inexpensive collection because there are little ongoing labor costs after development.

In a survey environment, TDE may be applied where the desired responses are numerical, or when responses can be linked to a numerical code, such as "yes" is "1" and "no" is "0." As in other applications, the respondent initiates the call to the collection computer which controls the flow of the interview. The computer asks questions in either a synthesized voice or from a file of

digitized phrases prerecorded by a human speaker. After each question, the respondent keys the answer. The computer also repeats each entry for verification directly with the respondent, and an acknowledgement is required, such as "1" equals "correct."

TDE offers many advantages over other collection methods. In repetitive surveys, the respondent retains a single form for monthly or quarterly calls, reducing the costs of both postage and the labor involved in mail handling, both outgoing and incoming. Costs for data entry and data verification are eliminated. Most importantly, the uncertainty about sample status is minimized. The status of the sample can be assessed through analysis of the received calls versus the list of active TDE respondents. Informed judgments can be made about the timing and extent of the nonresponse workload. No time is lost while survey forms are in the mail or waiting for data entry. This is especially important for time-critical surveys.

TDE also offers convenience for the respondent. The computer is always available to accept the calls. For busy respondents who are frequently out of the office or away from home, in meetings or traveling, this feature may be preferable to scheduling calls in advance and risking interruptions and repeated callbacks. TDE reporting may require less time than CATI.



TDE has some limitations that should be carefully addressed in each survey environment. First, not all respondents have touchtone phones. Thus, implementation of TDE would likely be in combination with other collection modes, adding to the complexity of survey management. As with mail collection, the respondent also may need to be reminded to call in, although a simple advance notice postcard has proven very successful when properly timed.

#### Examples of Current Use

The only known survey application of TDE is the Current Employment Statistics (CES) survey at the Bureau of Labor Statistics (BLS). The CES program covers over 300,000 non-farm business establishments monthly. The, data items are few, essentially employment, hours paid, and earnings, and the CES is conducted by mail in conjunction with each state, the District of Columbia, Puerto Rico, and the Virgin islands. collection of CES data is time critical. Preliminary estimates are published after 2 weeks of collection. Thus, the time lost due to the variability of the mails has A severe impact on response rates.

Initial experiments were done using CATI. Large scale tests of CATI collection, involving 13 states and over 5000 respondents monthly, successfully showed the ability to collect data from the vast majority of respondents in time for the first publication. More than half the CATI sample was drawn from chronically late

20

respondents. Response rates are routinely 85 percent versus 50 percent for mail.

The higher costs of CATI stimulated interest in TDE self-response. The results of small scale tests in 4 states suggest that TDE can retain high response rates over a sustained period. Calls average less than 2 minutes, and about 25 percent of respondents are given short reminder calls just before the collection deadline. BLS is expanding TDE use to over 15 states during 1990.

Procedurally, the combination of advance notice postcards, timed to arrive during the reference period, and short nonresponse calls provide a strong, inexpensive collection process. TDE respondents receive a package of materials that explain the new collection method, how it differs from mail and telephone collection. First-time TDE users are requested to call the computer on a test basis using special codes before they are asked to submit real data. The machine readable data are uploaded to mainframes for further editing and reconciliation.

The respondents chosen for the first TDE tests were drawn from those under CATI collection. In this way the higher costs of CATI can be offset by savings from TDE. Other TDE tests targeted mail respondents who generally reported on time.

The widespread use of touchtone systems has spawned an industry-wide working group to standardize features (e.g., the key on the telephone) to simplify user access.

Definition -- Voice Recognition Entry

Voice Recognition Entry (VRE) is Just developing as a technology.

The characteristics of VRE are essentially the same as TDE. The respondent initiates the call to the computer, but instead of using the touchtone keypad, the respondent speaks to answer, in this application the spoken digits 0 through 9 and "yes" and "no." Both "oh" and "zero" are recognized.

There are two essential features for VRE systems. First, they should provide speaker independent recognition, meaning that almost any voice can be recognized without any "training" of the system. Some systems require extensive training of the software for each voice. While this is used in some office dictation systems, it is probably impractical for survey operations. Also, systems should provide for rapid entry of responses using continuous or connected digits. These features are commercially available for both microcomputers and minicomputer applications.

VRE also has limitations in application. First, VRE is only

applicable to respondents with access to a phone, a small but

21

unavoidable problem. Recognition accuracy is the primary determinant of respondent acceptance. The system in use at the Bureau of Labor Statistics was designed using speech profiles drawn from the mid-western states. Dialects from other regions may reduce the accuracy of the recognition leading to respondent frustration and low acceptance. Early test results suggest that recognition remains high in Maine, the home of a very difficult dialect for the speech interpreting algorithms. More testing is planned to decide the limits of current technology. Improving, recognition accuracy is the primary objective of the companies involved in speech research and development.

Development of VRE is presently limited because there are few current applications to provide advance training and public

acceptance. Early results suggest that respondents familiar with TDE and VRE prefer the later as more "natural." This finding points out the differences in questionnaire design. TDE questions ask respondents to "enter" data, whereas VRE respondents are asked questions in a manner similar to CATI because the responses are spoken. Recently, experiments using voice recognition have begun to appear, conveniently providing training for future survey respondents. Also, the similarities between TDE and VRE may minimize acceptance problems.

Both TDE and VRE applications at BLS use short questionnaires.

These techniques may limit the length of the survey, but this requires testing. They provide convenience and low costs, but respondents may balk at long lists of questions and the current limitation on the range of allowable answers to numbers and a few words. VRE offers a variety of interesting research problems in speech recognition and natural language understanding. These systems have not yet come into widespread use.

## Examples of Current Use

The BLS is now conducting tests of voice recognition in the CES survey. The procedures will parallel those used for TDE and will assess the effectiveness of VRE for the entire U.S. population. They will examine any limitations involving multiple telephone systems, geographic distances, and respondents, acceptance. Acceptance by respondents has been high.

## Potential Uses

These computer assisted self-response methods have wide potential applications. Ideal surveys are repetitive, short and numerical, especially if the data are entered into a computer before the call is made.

the population. Since eligibility is usually determined by very few criteria, a mailed form could direct the respondent to call in the answers to one or two questions to a central computer. After entering the unique identification number, the respondent would answer these questions. Then the survey manager would use the machine readable file for nonresponse follow-up and subsequent sampling.

BLS is considering TDE for pilot tests of survey supplements and other special one-time surveys to reduce costs and add valuable control, to augment or replace the traditional mail process, and to gain experience in the design and use TDE systems. The logical extension of existing TDE and VRE technology is the linking of them into a single system. For example, respondents call the system which then asks the respondent to respond by touchtone. If the tone is not recognized, the respondent is automatically switched to a VRE component. A third feature would be available to record changes in the respondent's attributes (e.g., name or address), or to record open-ended responses for later transcription -- voice mail.

Self-response methods are not limited to survey applications.



Any ongoing project that collects cost, workload or other management data could use self-response methods for inexpensive collection. For example, a large copier company uses TDE for collecting billing information. Equipment renters are required to call in the monthly usage levels by entering copier usage as touchtone data. The computer then generates a bill in response to the touchtone entry. Also, the U.S. Postal service uses TDE to link callers to prerecorded tapes covering the most frequently asked questions. The BLS will begin using similar technology to answer routine inquiries for economic information.

#### Future

Voice technology is still being developed. "The NIST report argues that the most natural mode of data collection is not paper or keyboards, but speech" (William Nicholls, 1989). Recorded voices are currently being used in some surveys. Speech technology includes voice simulation which is useful today in TDE applications. While numerical and very limited vocabulary are being used in data collection, it will be some time before automated speech systems will

be used to recognize free-form human speech in. a telephone interview  
or in a personal interview setting.

#### Summary

Some items to consider when deciding between data collection

methods are as follows:

23

1. CATI offers cost saving over the personal interview setting  
and would be useful for a large complex survey environment.  
However, it misses people without telephones.
2. CAPI retains the benefits of a personal interview setting  
where response rate is important, and does not require a  
telephone.

3. TDE is cheaper than CATI, but cannot handle the complex survey, and respondent acceptance is a concern.
  
4. PDE is typically used in an establishment survey. It does not require a separate-key entry stage, but requires respondents to have access to a terminal, typically a PC.
  
- 5 . VRE will see only specialized application in the medium term.

Whichever technique is selected, the integration of the electronic data collection method into computer based survey system should be considered. For example, address; labels and other administrative items must be created from the sample database, then the interview proceeds, editing is done, and the resulting data are fed into the analysis or summary system.

Also, the decision maker should consider whether to use a single or mixed mode of data collection. Two examples of mixed modes are the "Census" integrated CATI/CAPI design, or the "BLS" integrated TDE/CATI

design. William Nicholls comments that "In the long run, the best data collection strategy for establishment surveys may prove to be a readiness to accept whatever combination of methods the respondent finds most convenient." The creation of new technologies and improvements to existing technologies will continue to have an effect on data collection methodology.

#### IV. Methodological Issues

##### IV.A. Human-Machine Interfaces

###### Introduction

The design of the interface between a person and a computer can decide the success or failure of the interaction. Although the situation is improving, there is generally too little attention paid

to the effect of interface design on user performance. Interface design is often not considered until the last stages of software development when the total design has already been "locked-in."

Automated surveys will involve people with widely differing abilities using machines ranging from manual data-entry devices to powerful computers. Interface issues will reflect this diversity in people and machines. There is no one interface that will satisfy all needs. The relative importance of a given interface issue will depend entirely on the context of person-machine environment. Nonetheless, there are some guiding principles of user interface design.

CASIC benefits from consideration of user-related factors in interactive systems, interaction styles, interaction devices, response time considerations, system messages, printed manuals, online help, tutorials, and development styles. Many of these topics involve detailed consideration of how to present the computer power to the user. For example, interaction styles can be broken down into command languages that the user must learn before using the computer, menus that guide the user through the necessary procedures, and the direct

manipulation of objects whose icon representation appears on the screen. similarly, interaction devices can take on many forms -- keyboards, function keys, pointing devices, speech recognition, displays, printers, etc.

Techniques for automated information collection include CATI CAPI, computer assisted self-response surveys, and prepared data submission on tape. Except for tape submission, these techniques involve user interface design considerations. All must be successfully used with little or no training. The user interface must be "self-evident." Error recovery is important. The user must be protected from making errors wherever possible. When it is possible for the user to err, the recovery procedures must be positive, helpful, and easy to follow.

#### User of the Interface

It is essential to determine who the user of the interface will be before designing the interface. In automated statistical surveys, a user may be a well trained and highly motivated survey

professional. At the other end of the range, the user may be a first-time or only grudgingly cooperative survey respondent. Even within somewhat narrow user populations, there will be differences among users that can affect the usefulness of the interface. It may not even be possible to design an interface that perfectly suits a single user because the user is subject to changes over time due to personal factors, new experiences, and changing needs. A user-interface design team should include an applied psychologist to help determine the psychological profile and needs of the user. The personality, training, and experience of the potential users are large factors in determining the most appropriate interaction style or styles for the user interface.

Interaction Styles

The choice of interaction style is also affected by the hardware to be used in the survey. Survey techniques that make use of computers with standard input/output devices can use command languages, menus or direct manipulation. Command languages are used to interact directly with the operating system of the computer. They allow a wide range of system functions -- storage, deletion, copying and printing of files -- to be done. The cost is a steep learning curve to master the commands. Command languages, while hard to learn, are also easy to forget. They can be intimidating to novice users who realize that information can be lost or damaged by poorly chosen commands. On the other hand, a person familiar with command languages can work rapidly and effectively. For some people, mastery of a command language is a source of pride which provides a sense of satisfaction and motivation for good job performance.

Menu selection represents another approach to interaction style.

Menus present the user with a set of only those choices that are appropriate at a given time. The choices are often numbered or lettered so the user can choose by entering the appropriate number or letter from a keypad or keyboard. Sometimes the choices are keyed to



the first letter of the line containing the choice. Then, the designer must be sure to avoid duplicate use of the starting letters. Some menus use pointing devices such as cursor keys, a trackball, a joystick, or a mouse to highlight choices. The user moves the pointing device to make a choice, then pushes a button to make the selection. Also, menus may offer only single-line choices. For example, a menu may ask for confirmation of a request by entry of y (for yes) or n (for no).

Menus are often organized hierarchically in graphs - data structures used to represent relationships among objects. Family trees are a form of graph that show the relationships of a person to other family members. Airline route maps are graphs that show paths the airline follows in flying between locations. With menus, the user is essentially "flying" by making selections from the

graph of menus (the technical term is "walking"). Selection of one item from a menu takes the user on a different path through the graph than does selection of another item. Graph structures can ease the design problem for complex user interfaces, but also can lead to user confusion. The user must be able to maintain a sense of location in relation to previous choices made. The user also must be given easy access to "escape hatches" if an unwanted path (undesired choices) has been walked on the graph. CATI and CAPI designs rely heavily on complex branching structures to control the interview. The menus and list of allowable responses must be clear, exhaustive and enable the interviewer to retain effective control.

Direct manipulation (DM) interfaces offer a third approach to interaction style. In DM, the user is given the impression of directly interacting with the objects of interest. As an example of a DM interface, consider a modern word-processing system. The screen representation of the document is made to be as close to the appearance of the finished document as possible. This is sometimes called WYSIWYG, (pronounced "whizzi-wig"), for "What You See Is What You Get." The user operates directly on the screen representation of

the document and immediately sees the results of the operation. Many commercially available graphical interfaces show how far DM can go toward helping the user. A mouse is typically used as the pointing device to objects on the screen. A typical screen object is an icon that symbolically represents the object. To delete a file, for instance, the user simply points to the file name and "drags" it over to a trashcan icon.

Menu selection and direct manipulation are important user interface techniques in situations that involve novice users with little opportunity for training. Although the interfaces must accommodate novice users, they also must be flexible enough to avoid frustrating more experienced users. Direct manipulation can accommodate novice and experienced users equally. Menu systems should allow experienced users to "select ahead" or to revert to a command language style of interaction.

Survey techniques that do not use more-or-less standard computers will raise unique interface issues. Alphabetic input, such as name entry, in telephone keypad-entry systems raises the question of letter assignment to keys that have multiple letters on them. Disambiguation

may be possible when the entries can be compared to a fixed list of permissible entries.

Speech recognition and synthesis devices have the potential for radically changing the preferred interaction style in user interfaces.

Although speaker-independent recognition of free-form spoken natural language is still in the future, rapid technological advances are being made in the ability to recognize automatically a subset of articulated words. Advances are also being made in the ability to synthesize natural-sounding speech under computer

control. The best form of human-machine interfaces in any give situation or for any specialized group of users is still a research question. This can lead to degradation of the quality of the survey due to user errors and frustration.

Some survey techniques are already speech based. In CATI and CAPI, the user interacts with a speaking and listening person who is visually and manually interacting with a computer. The person conducting the survey uses common sense to interact with the respondent. Although there are substantial efforts to imbue a computer with common sense, practical use of this research remains in the future. Thus, the effective replacement of the human interviewer by a computer also remains in the future.

#### Error Avoidance and Recovery

Whenever possible, interfaces should be designed so that errors are not possible. The nature of potential errors in a given interface must be thoroughly understood to lessen the probability of their occurrence and the cost of recovering from them. When a particular sequence of operations is necessary to do a complex operation, the interface should be designed to combine the entire sequence into a single operation. This will reduce the number of operations required of the user (who probably thinks of the sequence as one operation

anyway). All displays must have consistent layouts so the user does not have to spend time and mental energy scanning the screen for information.

The interaction style can have a profound effect on errors.

Properly designed menu systems can reduce errors by simply not offering poor choices. Choices offered must be clearly labelled. The consequences of a choice must be shown before the choice is made.

There must be consistency between menus. For example, a choice common to all menus (such as Cancel Menu), must appear in the same place in each menu and must have the same consequence (such as reversion to the previous menu).

Error messages should be designed to help the user. The messages should be specific, positive in tone, and constructive. They should tell the user what can be done to correct the error. Whenever an error is made, the user must have a clear and easily followed path to recovery. This not only reduces the seriousness of the consequences of the error, but increases the user's confidence even in the face of a few errors.

Adequate training can help to reduce errors and increase respondent acceptance. Certainly, respondents should be trained before using the system. Good training can be reinforced by providing on-line or telephone-accessible help and on-line tutorials. on-line or telephone-accessible help gives the user an

28

immediate reminder about proper operation of the system. On-line tutorials allow the user to review the correct procedures.

#### Design of Automated Form

In general, automated forms should not be automated versions of the manual forms they replace. They should be designed from scratch to consider to make use of opportunities and limitations introduced by automation. Sometimes, it might be appropriate to maintain the same "look and feel" between a manual form and its automated counterpart.

For instance, user training might be reduced by minimizing changes.

In these cases, the form designers should compare the benefits of staying with the old form with the costs of designing a new form.

Automation provides opportunities for higher productivity, lower errors, and greater user satisfaction over manual methods. Repetitive information can be automatically filled in from one form to another. Automatic editing for internal consistency and logical consistency should help to lower error rates. Automated forms also can provide on-line help and tutorials for the user.

Automated forms need not even look like paper forms. The user can be led through an interactive dialogue while the computer does the data formatting. Form fill-in is just one interactive style. Menu selection has already been mentioned as another style. Form designers should consider using hypertext, a recent development in interactive systems which provides a browsing environment. For example, the reader can display a definition simply by pointing at a word or phrase with a mouse. Hypertext would allow non-linear traversal of forms, as appropriate for the data being filled in. For example, in surveying



for medical information, gender data can be used to steer the user around inappropriate survey questions.

Form designers should have a repertoire of techniques for designing and testing forms. Expert systems might be developed to help in form design and interaction design. Effort placed in designing expert systems would pay off handsomely in easing individual design tasks. Such systems also should produce forms that are more consistent and complete than forms produced in a paper environment.

#### Quality Measures

It is critically important to test user interfaces before presenting them to the users. Professor Ben Shneiderman of the University of Maryland has identified five goals that lend themselves to precise measurement:

1. Time to learn - how long does a typical user take to learn to use the system?
  
2. Speed of performance - how long does it take to carry out a benchmark set of tasks?
  
3. Error rate - how many and what kinds of errors are made by typical users?
  
4. Subjective satisfaction - how much do users like using the system?
  
5. Retention over time - how well do users maintain their knowledge?

It is not enough to guess how well a system meets these quality

measures. It is essential to test the system. A testing laboratory is essential for any significant design work. Design groups may build in-house laboratories, or may seek help from existing laboratories.

It often happens that persons who are skilled in computer programming, data collection techniques, or statistical methods are not fully aware of the skills and deficiencies of the user population. It is not a good idea to concentrate the entire design effort in the hands of task specialists. The human factors role must be an integral part of every design team. Large teams might include psychologists, sociologists, and other human factors specialists. Smaller teams should at least assign one team member the role of human factors specialist. If nothing else, this person can play "devil's advocate" to be sure the appropriate questions are raised.

Data about user performance under current conditions must be collected before beginning new systems. It will not be possible to determine the relative quality of a new system unless quantitative measures of the quality of the old system are available. The first task of the design team must be to develop guidelines for the design.

Such items as menu selection formats, terminology, screen layout, data entry formats, error messages and recovery procedures, on-line help,

and training should be considered and decided upon before any other significant design work is begun.

Rapid prototyping is a powerful technique which allows iterative convergence to a design. Partial system implementations are made quickly, presented to potential users, and tested. Further development is based on these interim tests. Because each step in the development cycle is small, and tested incrementally, only small corrections in direction are needed at each step. Conceptual errors are quickly uncovered and are easy to correct. Rapid prototyping methods contrast sharply with the more conventional "waterfall" design methodology. The waterfall method requires detailed up-front specification of the design, with a

full-blown design flowing down to a full-blown implementation. While

this method may be appropriate in situations where the goal is clearly understood at the start, it has the disadvantage that changes made in any phase of the design tend to be large and expensive. This usually discourages change and leads to Acceptance of a lower, quality product or total abandonment of the design. A disadvantage of rapid prototyping is that formal specifications and documentation may never get produced in the flush of excitement over the rapidly evolving (and working) system. The waterfall methodology is appropriate as the final phase of a rapid prototype design. Because rapid prototyping quickly produces a working model and deep understanding of goals and tradeoffs, waterfaling can be effectively used to provide the missing rigor and discipline.

Evaluation must continue. even after a design has been completed and fielded. on-line suggestion boxes and trouble reports, designed right into the survey forms, provide easy channels of communication between the user and the designers. A user who suggests improvements or reports trouble should receive prompt responses and fixes. Large surveys might consider the use of a commercial bulletin board system as the communications medium for problems, suggestions, and fixes.

## IV.B. Software Development

### Introduction

There are two types of software that will be discussed in this section: software that helps in the creation of a survey questionnaire and software that makes up the actual programming code to execute the survey questionnaire in the field. This distinction is directly analogous to the usual notion of a highlevel programming language (e.g., FORTRAN, COBOL) in which you describe the problem in terms that humans can understand. This high-level description is then passed to a compiler that translates the description into an application program the computer can understand. For convenience, refer to the survey creation software as the survey definition process and to the use of the resulting application program as the survey application process.

Most of the discussion will relate to the creation software.

Historically, software development for automated field data collection began with a mainframe application for CATI. As hardware technology progressed, CATI was moved first to a minicomputer and then to a microcomputer. The CAPI application became possible with the development of the "light weight" portable microcomputer. Software to produce an automated questionnaire is perhaps the most important and potentially the most costly ingredient in the automated field data collection equation. Ideally, such software should be available off-the-shelf . Although there have been several attempts to develop such software, success has been limited.

To date, the development of automated questionnaire software has been done in one of two ways. The questionnaires are custom programmed using one of a variety of general programming languages (e.g., Pascal, C, FORTRAN), or they are custom programmed using a specialized CAPI/CATI programming language.

The specialized languages generally provide a means to describe a

variety of attributes: the question text; the answer text; the type of answer expected, (e.g., single, multiple, fill-in, free text) ; question paths (e.g., simple -- go to next question in order or complex -- based on the answers to previous questions, or some related calculation); response editing (e.g., restrictions to specific values or - range of values) ; and in some instances, screen layout design. In either case, the development of an automated questionnaire usually has required the skill of a computer programmer.

## Flexibility

There are several issues that need to be considered in the development or purchase of existing software for automating field data collection of survey questionnaires. Among these considerations is the level of flexibility needed. Flexibility is defined in terms of the amount of control the automated questionnaire exercises over the



conduct of the survey and in terms of the features available to design an automated questionnaire.

With respect to the control, consideration must be given to the extent the automated questionnaire will allow the interviewer or respondent to exercise control over the conduct of the interview. That is, should the person controlling the interview have the same control as in a paper-and-pencil conducted survey; total freedom to roam anywhere in the Questionnaire and change questionnaire answers at anytime or should the automated questionnaire be designed to limit the person collecting the data to a specific process and skip patterns or some level in-between? If so, what is that level? The answer to these questions is critical because the software selected, particularly if it is a specialized package, might not have the specific capabilities needed to implement the desired design. The design of the questionnaire software also will be affected dramatically by the level of flexibility chosen.

With respect to software flexibility, there are several capabilities that should be considered. These capabilities are:

1. The question types: open ended, closed ended, single value, multiple values.
  
2. Case management: administration of each questionnaire, e.g., status of completion, restart .incomplete questionnaire.
  
3. Back-up: ability to back-up to any question in the survey and change an answer, with the system thereafter automatically following the skip patterns implied by the changed answer.
  
4. Editing: ability to perform edits such as consistency, range, and specific value or values.
  
5. Screen manipulation: ability to create any screen design desired.
  
6. Comments: ability for person recording answers to record comments associated with any question.

7. Skip patterns: simple and complex, e.g., skip based on answers to previous questions or some arithmetic calculation.
  8. Context sensitive help: ability to get help based on place in survey.
  9. Rostering: ability to handle household member enumeration, identification, and skip patterns based on the individuals.
  10. output format: form collected data is stored, e.g., a flat file.
- ii. Accessibility of collected data: how easy is it to access the data, e.g., quality control.

12. Coding: ability to code collected data automatically or manually.
  
13. Authoring system: ability to create questionnaire and software to execute the survey questionnaire (program code) simultaneously with no computer programming skills.
  
14. Output reporting: reports about the functioning of the data collection process and about the actual data collected.

This list of features is not inclusive, but does contain the most important features determining the level of flexibility.

Range

There are several additional factors that are important to the decision of level of flexibility and software design. These factors are the size and complexity of the survey questionnaire and the period

between major changes in the questionnaire or the preparation of an entirely new questionnaire. Complexity is defined by the number of different question types, complexity of skip patterns, and need for Fostering. Size and complexity are directly proportional to software development time. The shorter the period between major software developments, the greater the requirement for a user-friendly authoring system. An authoring system significantly decreases development time and decreases computer programmer dependency. The size of the questionnaire also may impact the hardware and software requirements. Several software packages have certain restrictions that may be affected by the size of the application.

#### Automated Forms Design

Unlike CAPI and CATI software, there are many off-the-shelf software packages that can produce automated forms for computer assisted data entry. Many CAPI and CATI specialized software packages

also can be used for this function.

## Training

The amount and type of training required to use selected survey questionnaire development software is dependent upon the level of user-friendliness of the software. For example, programming the questionnaire in Pascal would require considerably more skill and therefore more training than programming the questionnaire using an authoring system. Usually, it is necessary to have a skilled computer programmer working with the survey questionnaire designer in order to use the current software. Under these circumstances the questionnaire is most likely to be a pencil-and-paper questionnaire programmed for the computer rather than one designed for the computer. Computerized questionnaires will improve in quality as their designers come to understand and use the environment provided by the computer.

Software documentation for the specific survey questionnaire should be complete enough to insure easy revision of the questionnaire by someone other than the original author. For the general

programming languages there are many software packages available to help in such documentation. The liberal use of comments in the computer programming code also is a good way of providing additional documentation.

#### IV.C. Data Collection Programs

##### Introduction

When producing a survey, several factors will affect the selection of a data collection method. The three primary factors are cost of resources, the time available to collect, edit, and summarize the data, and the desired quality. Because it is unusual to have all three in abundance, trade-offs must be considered.

Several other important factors relate to the design and operation of the survey, and will affect the cost, timing, and quality

factors. First, the survey may be one-time or ongoing. A one-time survey may want to maximize quality for a fixed cost, where an ongoing survey - may want to maximize quality for a minimized cost. With ongoing surveys automated capabilities can evolve over extended periods thereby spreading out the costs. The second factor is the target population, and whether it is a household or an establishment. The chance of finding PC's in establishments is greater than in households, although not all households have telephones. The third factor is the operational nature of the survey, that is whether the setup should be centralized or decentralized, and whether the PC's would be networked. Lastly, the sample size and complexity of the questionnaire is relevant.

The remaining nine factors relate to the characteristics of the technology used to collect data.

1. The Speed at which data may be entered is determined by the technology's hardware (such as XT, AT, or 386 PC's, disk speeds, and phone lines) and software (the complexity of the questionnaire and therefore the length of the program).



2. The Size of the machine can refer to its weight or ungainliness (which is important in situations where it must be moved around) or its available memory (which limits the amount of data and the complexity of the program that can be stored on the machine).
  
3. The portability of a computer's software is important in situations where data collection is carried out on different computer systems.
  
4. The Type of Display selected may be based on environmental factors (where conditions are indoors and usually fixed, or outdoors and variable therefore screen color is important), and on the complexity of the questionnaire (and therefore screen size).

5 . The Mode of Data Entry varies from keyboard, to push button phone, to voice data entry.

6 . Data verification is based on the importance of quality, the complexity of the data, and other factors as hardware speed and available memory.

7. The Database Generation refers to the way in which the data is brought together and integrated with the rest of the survey system. This may mean using telecommunications, or simple computer tasks.

8. The Hardware selected is based on cost, amount of time available, data quality desired, power of the machine, amount of memory, and other available features.

9. Training is important in any survey, and the amount of time available and the background of the staff dictates the

technology chosen.

The priorities of these factors and the relationships between them help to decide which data collection strategy to use. A discussion of these factors with regards to CATI, CAPI, and other methods, follows.

CATI

Introduction

In a CATI interview, the interviewer is helped by an interactive computer system. It provides data quickly and offers good reliability, but a substantial cost investment is required to purchase and set up the system. The cost investment may be greater than other electronic data collection techniques, but it saves money over face to face interviews, since data entry is combined with data collection. It also can be used for follow-up of nonrespondents or edit failures, or key in of mail questionnaires. It can be used in a household or

establishment survey with complex questionnaires (typically a new or infrequent survey where time series interruptions will not cause problems, and where sample size is large, or small and used over a longer period). It can be operated in a centralized or decentralized manner, but it requires the respondent to have a telephone.

Hardware: The first generation consisted mostly of mainframe based systems, but the current generation consists of either multiuser minicomputer systems, or distributed systems over a PC local area network (LAN). The minicomputers are often UNIX-based and

used mainly in large centralized facilities that require greater resources to pay for specialized support staff. The PC's are mostly DOS-based and are used in multi-location facilities. An added benefit of PC's (even in large facilities) is that many clusters of networks can be used, and PC's can be added one at a time (lower initial cost).

Speed: With minicomputers, the speed between questions could slow as the number of interview stations increases, or if another computer intensive program is run. With PC's on a LAN, the speed between interviews could slow as more stations are added to the network. Eventually, faster computers will solve this problem.

size: The organization of the system (centralized or decentralized) and the hardware (minicomputers or PC's) will affect size requirements. The system can range from a single stand-alone PC to 100 or more workstations on a mainframe system. The PC's and minicomputers usually have from 5 to 60 networked workstations.

Portability: The software should run on multiple hardware platforms with different operating systems. It should be written in a portable language and use common user interface standards. Today, software costs are increasing while hardware costs are decreasing. Portable software should provide a cost savings across different hardware platforms.

Displays: The use of color can aid the interviewer, but the Color Graphics Adaptor (CGA) standard is not clear enough for use over a long time. Either the non-composite monochrome, the higher resolution Extended Graphics Adaptor (EGA), or the very high resolution Video Graphics Array (VGA) standard should be used. However, EGA and VGA are more expensive.

Data Entry: screens can be item based, screen based, form based, or a combination of these. Movement between items can be forward only, or forward and backward. Most systems have question skipping and branching capabilities, interviewer notes can be added, and the interviewer can resume at the point where the previous session ended.

Data Verification: The data quality is improved by incorporating longitudinal (historical) editing, arithmetic calculations, range, and consistency checks.

Database Generation: Outputs consist of an audit trail and response data. Often numeric and open ended data is stored separately, then linked by respondent number. Some systems include cross-

tabulation capabilities, and the ability to generate accurate and timely reports is a benefit.

38

Training: one benefit is that centralized supervision and monitoring is available (on-line and audio-visual). It helps the supervisor identify interviewers who need more training.

CAPI

Introduction

In CAPI, the equipment is less expensive than CATI, but travel costs are higher. It requires the same amount of time as personal interviews, but data quality is improved and the separate data entry step is deleted. One advantage of the personal interview setting is that it causes higher response rates.

Hardware: The following criteria can be used to evaluate potential portable computers: interview duration and complexity, memory capacity, weight, power source, and duration, screen size and legibility, disk type and capacity, speed, serviceability (important because service centers might not be locally available), portability, durability, price, ease of use and software compatibility.

Speed: The speed depends on the computer hardware and complexity of the questionnaire.

Size: A larger portable computer would be needed to put a complex questionnaire in 2 languages. Even a small portable computer is not necessarily portable as many have complained that they are too heavy to carry around for very long. Electrical outlets are not always available. The battery power required for additional memory and for disk drives can add substantially to the weight requirements.

Although small portable computers can be used on a table top or in one's lap, interviews conducted on the doorstep require handheld computers. That technology is coming but has yet to arrive for



general use. A smaller portable computer, or one with a different keyboard would be needed for this environment.

Portability: As in CATI, the questionnaire writing software is often portable from one type of hardware to another.

Displays: Different portable computers have different size screens with various readability factors. The various lighting conditions that would be met in the field is also a factor. For example, a "back light" screen is required for dim lighting conditions. If the interviews are conducted outdoors, glare reflection is a problem.

Data Entry: often the software that was designed for CATI is also used for CAPI. It provides forward and backward movement, and incorporates skipping and branching between questions.

Data Verification: Similar to CATI, improved data quality results from reduced clerical and machine activities, and being able to incorporate various editing techniques.

Database Generation: Data output can be consolidated more rapidly due to reduced clerical and machine activities. Data transmission options are mail, courier, or phone lines. Data security and the quality of phone lines may be a factor against using phone lines.

Training: Basic interview skills are considered very important (even more so than computer knowledge) . With this assumption, training should on the computer and questionnaire details. Training materials can include a tutorial (helps coordinate the different learning rates), self study materials, and hands on practice with interviews. Good software and manuals are also important.

Data collection using TDE requires the respondent to have a touchtone telephone, and a dedicated computer with a multiple phone line capability at the other end. one benefit to the respondent is the convenience to call in at any time.

Existing TDE systems limit editing primarily because of limits on hardware capacity, lack of visual clues and restriction to push buttons on the telephone. However, the computer can synthesize the answer and play it back to the respondent thereby providing the opportunity to verify or correct the answer. TDE offers lower cost than CATI (less labor and mail costs with key-entry costs born by the respondent), and the data quality is good. TDE has been able to retain very high response rates over long periods when coupled with appropriate nonresponse prompting.

VRE again requires only a telephone and carries a cost profile similar to TDE.

Surveys which use PDE require the respondent to have access to a microcomputer. Data can be entered using the keyboard or a file containing the data can be imported. Displays are typically an

electronic image of the form on the screen. Error checking and other edits can be included, after which the data is transmitted back to the required agency where it is combined with other data. Computer - security issues are important here. Integrity checks to make sure the data received is the same as the data sent must be part of the system. Appropriate manuals and other training materials including on-line help should be provided. This type of data collection would be worthwhile in an establishment survey where respondents report data monthly, quarterly, or over a given period.

#### IV.D. System Interfaces for Data Conversion

##### Introduction

Automated submission of data has the benefit of reducing reporting errors because a keying step can be eliminated.

Traditionally, respondents entered data onto paper forms which were mailed to central site where they were keyed into a computer system. With automated data submissions, intermediate keying steps can be eliminated.

Automated data transmission requires hardware and software compatibility between the respondent site and the Federal site. In recent years the number and types of software and hardware options have greatly multiplied into the current myriad of products and technologies on the market. Due to these developments, Federal agencies are often looking at heterogeneous sources for data transmission.

Federal agencies conduct many surveys with many types of respondents. These data sources, such as state and local governments and businesses, will increasingly have capabilities for reporting data in an automated way. Many now have personal computers (PC's) while others have only mainframes available. Complexity arises as Federal agencies, looking at a mix of hardware and software technologies available at respondent sites, must select the best way to collect

data from these heterogeneous sources.

### Planning for system Interfaces

Managers of data collection projects can expect interface problems, but these problems can be minimized by good planning. Knowledge about the availability of communications capability, hardware, and software at respondent sites will aid managers in their planning for system interfaces for data collection.

### Communications Capability

Perhaps the most important issue for system interfaces is communications. Communications may be thought of as networking or as linking technologies together. With networking capability, data can be transmitted across telephone lines or special private line arrangements such as local area networks (LAN's). See the section on Networks Planning in this report for a discussion of networking

issues. A related issue is maintaining the confidentiality of data transmitted in such a manner. See the section on Computer Security in this report.

## Hardware

Hardware is needed at both the respondent site and the Federal site for data transfer. The type of hardware available at the respondent site will often decide what options the Federal survey managers will offer for submitting data. It may be necessary for the Federal site to have hardware for data conversion available, for example, hardware to read both 5 1/4 inch and 3 1/2 inch diskettes. Also, communications may need to be set up between hardware devices. The section on Hardware Planning in this report discusses these issues further. Three common types of hardware links are discussed below.

Mainframe to Mainframe: Data can be transmitted from one

mainframe to another via a communications network. Either the respondent or the Federal site can specify record layout and formatting instructions for data submission. Front-end processors can do data conversion before the data are sent to the host computer. Another option is submission of a computer tape in a specified format.

PC to PC: A link between two PC's can be established using a network system. Another way to transmit data from one PC to another is to mail the data on diskette. The record layout and diskette format would be agreed upon by the respondent and the Federal site. Because diskette sizes vary, the Federal site may need conversion hardware and software to read diskettes of different sites. Another option is to provide software on a diskette to the respondents.

Mainframe to PC: This type of hardware link combines the options described above. Again, a link can be established using a communications network. If the PC is at the respondent site, a diskette with software may be provided to set up the PC to send



data over to the mainframe in the-appropriate format.

## Software Compatibility

Although Federal survey managers usually cannot provide hardware to respondent sites to use for data transmission, they often can provide software for this purpose. If the respondent's software is used, the Federal site must have the same software or be able to convert the data to the correct format. Not only can different software products be incompatible, but two versions of the same software product can be incompatible. One version may have a higher level of functionality than the other. Again, there must be planning for document transfer. See the section on

Software Development in this report for more guidance on planning for software compatibility.

#### IV.E. Computer Security

##### Introduction

Computer security refers to the continued operation of computer applications at acceptable levels of risk to the organizations) being supported by the applications. Risk is usually measured in terms of potential loss, specifically losses that occur from:

1. Disclosure of information to unauthorized parties (i.e., loss of confidentiality),
2. Modification or other adverse actions that affect the expected quality of information (i.e., loss of integrity),  
and

3. Destruction or other adverse events that affect either the availability of the information when it is needed or the availability of the computer system to process that information (i.e., denial of service/loss of availability).

The types of losses described above can result from accidental and intentional events, as well as from natural hazards.

When estimating risk, it is important to consider direct losses (e.g., the cost to replace modified or destroyed information), as well as indirect losses (e.g., the inability of the organization to meet its mission which can lead to public embarrassment, congressional wrath, loss of lives, legal actions, competitive disadvantage, etc.).

After estimates of risk are derived, it is necessary to select and implement cost-effective safeguards (e.g., physical, administrative, technical, management) to reduce these risks to acceptable levels.

With respect to automated statistical surveys, the types of losses discussed above can occur during data entry from the respondent, during transmission of the survey information to the host

computer system, and within the host system. While the ideas discussed below are generally applicable to all of the survey types addressed in Section III of this report, this section will focus on surveys collected through or with the use of a computer where the following occurs:

1. Data entry using a terminal or computer system to collect the response information (i.e., not directly applicable to response information collected over the telephone). The data entry Process may "batch" the respondent's information for later transmission to the host computer for processing or may have the respondent connected

directly to the host system where the survey data is being captured in real-time (and may be processed in real-time).

2. Transmission of the response information over telecommunications lines/circuits, including future ISDN networks discussed above, and transmission on magnetic media (e.g., floppy disk) through public and private mail delivery services, and
  
3. Receipt and processing of the survey information by a host computer system.

#### Problem Areas

#### Data Entry

During the data entry process, the following issues need to be addressed with respect to computer security.

Identification and Authentication: Respondents and other users of computer systems that are used to collect survey information must be positively identified and authenticated to assure the validity of the survey and to hold users accountable for their accidental or

intentional actions. While passwords are still the most widely used method of authenticating the users claim of identity, other methods such as biometrics and smartcards can be used when increased protection is desired--usually at increased cost. Passwords can be effective for authentication when used in accordance with FIPS 112, Password Usage Standard.

Access Control: Access to information on computer systems should be strictly controlled so that users only have access to information they are authorized to see or change. Most commercial computer systems provide mechanisms that support this function. Systems that appear on the National Computer Security Center's Evaluated Products, List contain operating system level access controls that provide protection from unauthorized disclosure of information. Access controls are important on multi-user systems that are used to collect survey data in order to prevent the survey data from being intentionally or accidentally read, modified or destroyed.

Accountability: Unless computer systems contain mechanisms for recording and analyzing users, computer security relevant actions, it

will not be possible to hold users accountable for actions that cause computer-related losses. When users know that a computer system has an effective audit trail collection and processing mechanism, they are less likely to make mistakes or to attempt unauthorized access to information for fear of being caught. When survey data is collected on systems that provide

45

accountability mechanisms, it will be easier to determine if the survey data have been tampered with or have been disclosed to unauthorized users.

Confidentiality: Besides access controls discussed above for preventing survey data from being disclosed to unauthorized individuals, cryptography can be used to protect data while it is being stored in a computer system or on other magnetic media such as floppy disk or magnetic tape. FIPS 46, Data Encryption Standard (DES), defines the only government-wide standard for encrypting and

decrypting unclassified computer data. Since the DES has also been widely accepted by the commercial sector, there are many off-the-shelf-products that can be purchased for implementing DES cryptographic protection.

Integrity: During data entry, the integrity of survey data can be affected by entering false/inaccurate data or by modifying data already entered. Approaches for addressing these issues include;

1. Editing through the use of error detecting- or correcting software that determines reasonableness of input data with respect to any number of criteria such as character composition of data input, numerical bounds checks, data dependent checks on previously entered data, etc.
2. Access control (see above) that prevents unauthorized users from gaining access to the survey data
3. Cryptographic check sum as defined in FIPS 113, Data Authentication Standard that places a cryptographic "seal"



on the survey data for the purpose of detecting modification of the survey data from some initial state. This technique is useful when the survey data is stored-in computer memory or on magnetic media such as floppy disk or magnetic tape.

4. Accountability is the primary method for detecting modification to survey data by individuals who ARE AUTHORIZED (i.e., access controls do not apply) to access the data. While effective against both accidental and intentional modification, authorized users that intentionally modify data can subvert accountability controls if they have a high degree of technical knowledge about the computer system.
  
5. Software-engineering assurance techniques should be used in developing the data entry and other system

software to preclude errors from being introduced into the survey data through faulty software.

Restart/Backup/Recovery: It is necessary to plan for restart/backup/recovery activities whenever the data entry process is interrupted or the survey data is destroyed. Techniques such as maintaining backup files, permitting restart points in the data entry process, and planning for an alternative data entry processing capability are all directed at maintaining continuity in the data entry process.

#### Transmission

During transmission, the respondent's survey data are sent from the data survey system to the host system that will process the survey data. While authentication applies primarily to transmission of survey data through telecommunications networks, confidentiality and integrity techniques are applicable to telecommunications networks and mail delivery of magnetic media.

Authentication of host computers (e.g., the host computer of the data entry system) to the transmission network is required by and provided for most telecommunications networks to prevent unauthorized use of the network and to facilitate billing for network services. Sometimes, depending on the sensitivity of the survey data, it might be necessary to have the transmission network authenticate itself to the data entry host system before sending such data over the network. In this way, the data entry system can be sure that the survey information is being sent over the actual network rather than being given to an intruder that is spoofing the data entry system into giving the intruder the survey data. If the network lacks capability for authenticating itself, then techniques used for confidentiality and integrity described below may be considered as alternative methods of protection.

Confidentiality: The most common technique for preventing disclosure of information within transmission networks is to use cryptography. As discussed above, the DES is the only government-wide standard for encrypting and decrypting unclassified computer data.

Integrity: integrity with regard to transmission of survey data is the assurance that the survey data has not been altered, either accidentally or intentionally, during the transmission process. Cryptographic checksum techniques, as described above in the section on Data Entry Integrity, are effective in providing this protection.

Availability/Reliability of Network Services: Sometimes, particularly in real-time data collection and transmission, continuity of the transmission service can be very important to the

success of the survey activity. Discontinuities due to the unavailability of the network or some of its intermediate nodes or due to noise in the transmission lines can result in survey data being lost, erroneous, or delayed. This could be particularly annoying to a respondent that has to keep repeating the survey data entry process or is unnecessarily prompted for nonresponse. it is possible to

minimize such problems by using networks that provide error detecting/correcting procedures, dynamic routing around unavailable nodes, and other services that assure network availability and reliability.

#### Host Computer System

Computer security concerns at the host computer are similar to those at the data entry computer. The reader should refer back to these discussions to supplement the material contained in the corresponding areas below.

Identification and Authentication: All users of the host system, including the respondent data entry system, should be required to identify and authenticate themselves to the host system to assure the validity of the survey and to hold users accountable for their accidental or intentional actions. The same authentication techniques that were discussed for the data entry system apply to the host system.

Access Control: Access to information on the host systems should be strictly controlled so that users only have access to information they are authorized to see or change; in particular only authorized users should be permitted to access survey data on the host system.

Accountability: The host computer system should contain mechanisms for recording and analyzing users, computer security relevant actions in order to hold users accountable for actions that cause computer-related losses, particularly losses to the survey data.

Confidentiality: Besides access controls discussed above for preventing survey data from being read by unauthorized individuals, cryptography can be used to protect data while it is being stored in the host system or on other magnetic media such as a floppy disk or magnetic tape. As with the data entry system, the DES should be used for this purpose.

Integrity: on the host computer, the integrity of survey data can be affected-by entering false/inaccurate data during the data

entry process or by modifying data already entered. Approaches for addressing these issues include;

- editing
  
- access control
  
- Cryptographic check sum
  
- accountability
  
- software engineering/assurance techniques

Restart/Backup/Recovery: This is necessary when the host computer system's processing is interrupted or the survey data is destroyed.

Techniques such as maintaining backup files, permitting restart points in the host's processing sequence, and planning for an alternative

host processing capability are all directed at maintaining continuity in the host' s processing of the survey data.

49

#### IV. P. Hardware Planning

##### Introduction

Hardware issues are related to the type of Computer Assisted Statistical Survey system and the particular software to be used. The adage that says to "choose the software first and then the hardware" may be accurate if the software is already available. if software needs to be developed, however, it may be better to settle the main hardware issues first.

Hardware issues may be divided into the types of hardware needed and the criteria used for selecting products. We will explore these



issues for current and forthcoming products.

## Current Hardware - General Issues

There are certain hardware issues that arise no matter what the application. They may be categorized into ergonomic, performance, capacity, and cost issues. Ergonomic issues include keyboard layout and touch (a tactile response reduces input errors), screen visibility and readability, and adjustability of the computer.

Performance and capacity can usually be improved only at higher cost. However, if the hardware is optimally designed for the application in mind, no higher-cost may be incurred. For example, performance can be further divided into CPU and I/O speed. It may suffice to maximize only CPU or I/O speed. Software techniques also may be employed to improve performance: use a RAM disk for files that are frequently accessed, delay I/O operations until they can be more conveniently done, and use machine language routines for CPU intensive operations.

Core memory requirements are driven by software needs. The main question is whether the DOS RAM address space of 1 megabyte is sufficient or not. If it isn't, various options are available. By swapping pages of memory in and out as needed, the address space can be expanded. Note that extra memory is not usable without a software driver.

#### Respondent Data Entry

If respondents will be using their computers, try to find out as much as possible about the machines they have. Respondents may not have access to a personal computer (PC) even in a large company. For example, an accounting department may have a mainframe, but not a PC. IBM-compatible computers are the most common in the business world, but they may be earlier models. Software that respondents will be using should be tested on minimal

hardware configurations. Don't assume that respondents have extended or expanded, memory. A hard disk probably can be assumed.

The 5 1/4" diskettes are now the most common but the new 3 1/2" diskettes are coming into use. Capability for reading either type would be helpful. There is a compatibility problem between 5 1/4" high density (1.2 megabyte) disk drives and lower density drives. The latter cannot always read disks formatted by the former even at lower densities. Also, writing high density data on a lower density disk can corrupt the contents.

#### CATI

Computers for CATI must support interactive processing, e.g., a multi-user mini-computer or a PC network. Speed is the most important factor. The time from entry of one item to display of the next should be less than two seconds. To minimize data transfer problems, the system used for data entry should be the same or compatible with the one used for subsequent processing.

## CAPI

The main criteria for CAPI computers are screen readability, speed, and weight. Many portable computers are too heavy and awkward to carry around. A truly portable computer is necessary. While the lightest portable computers now weigh 4 to 7 pounds, the screens on these machines may not 'be good enough. Full-sized screens with good visibility require extra battery power that implies a total weight of about 10 lbs.

Screen visibility and readability have come a long way. Many types of screens are available: cathode-ray tube (CRT), liquid crystal display (LCD) backlit supertwist LCD, gas plasma, DC plasma, and electroluminescent display. Quality varies so much from vendor to vendor and within each type that it is difficult to make generalizations. Factors to judge include screen contrast, resolution, blur when scrolling, size, adjustability, and power consumption. The screen should be tested in environments that approximate actual interview conditions such as dim lighting.

Good performance is now available, but the cost can be high. The 3 1/2" diskettes are used on portable computers; their smaller size and harder cover make them preferable. The carrying case should protect the computer if it is dropped or banged. it also should have a government emblem or insignia to identify the interviewer.

The battery charge on a portable computer may last up to four hours, but some models portable battery packs that can be inserted as needed. Respondents might allow the use of their AC outlets.

A low battery indicator is helpful; nickel-cadmium batteries should not be recharged before power runs out. A car battery adapter is useful on the road.

Touchtone data entry (TDE) and voice recognition entry (VRE)

require special hardware cards and sufficiently powerful computers.

The current BLS TDE configuration uses a 286 PC with 640k RAM. One PC

can support many phone lines. BLS estimates that for a survey with

1.5 minute calls received during a 2 week collection period, one phone

line is needed for every 500 respondents so that during peak

collection periods respondents will get a busy signal less than 5% of

the time.

Facsimile (FAX) transmission requires a hardware card or a

separate FAX machine. There are machines that combine Fax, image

scanning, laser printing, and photocopying.

Telecommunication usually means analog transmission over phone

lines. Digital computers must have a way of sending and receiving

analog signals; the device that handles this is called a modem. The

main distinction between different modems is the speed of

transmission. Bits per second (also erroneously called baud) rates of

1200 and 2400 are the most common while 300 and 9600 are also used.

As a rule of thumb, about one byte is transmitted per 10 bits because of parity and stop bits. Therefore, sending and receiving a large data set can take a long time. Software should have error checking capabilities.

## Future Hardware

Besides general technology trends (smaller, faster, less expensive, more capable machines), a few specific observations can be made. International standards are taking on a new importance. Standards committees are no longer just reacting to de facto market standards but are taking the lead before products are developed. Compatibility and interconnectivity with other products are often as important as the capabilities of a product itself.

The future for portable computers is bright. Color screens and more memory and disk space are going into smaller and lighter machines. Handheld PC's are starting to appear. computers the size of today's miniature calculators are not far off. Cellular telephones will be combined with portable computers. Peripherals such as printers' are becoming more portable.

Electronic Data Interchange (EDI) is changing business practices by automating orders, invoices, etc. As this becomes more widespread, surveys could be designed to "piggyback" onto EDI

52

to take advantage of the systems already in place. Wide area computer networks with electronic mail are becoming more like public utilities.

Developments in digital- telecommunications (e.g., the Integrated Digital Services Network, or ISDN) will have many hardware implications -- see Network Planning. Modems will no longer be necessary because the entire path from computer to computer will be digital. Data transfer rates will be much faster.

Optical and optical-electronic technologies are dramatically increasing data storage capacities. High definition television (HDTV)



and digital video interactive (DVI) will intensify graphic applications. Improved optical character recognition (OCR) will help the transition from paper to completely electronic representation.

#### IV.G Networks

##### Introduction

The computer revolution has come upon us in a series of waves: the first computers transformed the speed of computation by several orders of magnitude; improved technology provided computer access to large organizations; personal computers provided computers to everyone; and the relative recent introduction of computer networks created the information community which has brought information to everyone.

Networks have made possible the development of information utilities that serve the entire spectrum of the human community, providing services from computer games to newspapers for anyone owning a personal computer. The pervasiveness of these information services enables survey information to be collected locally and transmitted directly to a center processing utility.

The Arpanet developed by the Department of Defense was the first widespread network to join researchers, system developers, and administrators into an information community. Although electronic mail or E-mail was the immediate gain from this network, the ability to transfer files of data, to access remote databases, and to use the computing services of a geographically remote computer showed the real value of a network.

Access to computer networks by the public has increased dramatically as the network cost for an individual has dropped to the cost of a local phone call. Some commercial services cost less than a monthly phone bill for unlimited access. A new network technology is about to transform our ability to use the distributed processing

systems available on a network by dramatically increasing the amount of data that can pass over these networks.

#### Data Collection

Networks will have a profound effect on data collection. They will provide the opportunity for close contact between the interviewer and the respondent. For example, CATI provides limited voice interaction over a telephone. Networks will provide visual and audio interaction with television or computer screens. They will enable the interviewer to display previously collected data to the respondent, and to use graphical diagrams and pictures to convey the conceptual background to questions. Moreover, it will provide the opportunity for more frequent updates to survey information that will match the data requirements rather than the economical constraints.

High-speed networks will put interviewers in closer contact with experts who can resolve troublesome issues while a survey is

being conducted. For example, CAPI interviewers do not have immediate access to their supervisors. With high bandwidth networks, the CAPI interviewer can contact a supervisor in much the same way as a CATI interviewer.

The net result should be greater interaction and reduced costs as the network bandwidth increases by an order of magnitude over the next decade.

## Background

There has been a separate, and independent, evolution of networks in this century for the transport of voice and data. The classical voice network was based on the telephone handset that converts speech into electrical signals which are transported over the local loop via a twisted pair of copper wires to a telephone system end-office.

Traditionally, the signalling involved has been analog (the transported signal varies continuously in time) and the communication

link established between two telephone handsets has been termed an analog voice transmission circuit.

The human ear is an extremely good filter, and has permitted analog voice circuits to be established in which the analog voice signal was noisy. A good ear and contextual information made it possible to understand the communication. As the separation between two handsets engaged in an analog voice communication link increased, the electrical signals required amplification for continued distribution. Such amplifiers are often called repeaters and they had the unfortunate characteristic of amplifying both the noise and the electrical voice signal being transmitted. Consequently, it was very difficult to remove specific noise components from the analog voice signal.

The analog telephone handset is connected to a local exchange or end-office. This is nothing more than a local switch that is in turn connected to a trunk exchange. This trunk exchange, in North America, is a five level hierarchical arrangement of switches for routing telephone calls. It forms a circuit switched network that is

connected to an international access exchange and provides the capability of global voice telephone communications. The network described here was still made up of twisted-wire copper pairs in the local loop and electromechanical switches that performed routing of the voice calls until 1966.

With the appearance of very large scale integrated (VLSI) technology -- the computer on a chip, the network switches evolved into electronic switching systems. The intelligence in the switches allowed the established transmission fabric to be rendered more cost-effective by simplifying maintenance through higher reliability features and better strategically planned network maintenance. As the employment of more sophisticated electronics

was accelerated in the switching matrix, the conversion of analog voice signals to purely digital signals led to the appearance of

dedicated digital networks. While the handset in most installations remains analog, the local switch to which the handset is attached performs an analog to digital conversion of the initial voice signal. From there the signal is entirely digital. The digital transmission networks that are dedicated to voice and data are called Integrated Digital Networks (IDNs). Standards have now emerged internationally using the guidelines of Consulting Committee for International Telephony and Telegraphy (CCITT) - The digital transmission systems are rapidly evolving toward IDNs that are interoperable and make use of the intelligence associated with each network switch that is digital because each such switch may be regarded as a generalized computer.

In making distinctions between data applications and voice applications using a modern IDN, networks used for data applications can be characterized according to the activities of the terminals on the network:

1. Start-stop terminals are used to generate interactive data traffic to and from the computer. This traffic-tends to be low speed with occasional bursts as the computer responds to

an interactive request for a specific file to be transferred.

2. Batch data transfers and data display image transfers that occur as bursts of data that can be placed on the network.
  
3. Continuous data traffic that is typically carried by circuit switched IDNs at data rates from 2.4 to 64 Kbits/sec (thousands of bits per second) . The data traffic within the network is often combined from separate low data rate bit streams, and interleaved into a single 64 Kbits/sec data channel for transmission across the network. A packet switched network decomposes a digital message into smaller chunks of bits (typically 1008 bits or 2000 bits) and routes these chunks, called packets, through the network from a source to a destination on an end-to-end basis.

#### Current Network Systems

The modern communications environment may be regarded as made up



of three basic functional blocks:

1. User terminals that support a human interface with the network. They allow a human to interact with

56

another user terminal or a computer connected via the network.

2. A communications network that is transparent to the user and provides conventional information transfer capabilities.

3. Information service centers that provide computing functions at the center.

Network systems breakdown conceptually into Local Area Networks

(LAN) and Wide Area Networks (WAN). The IEEE definition for a LAN is

a "data communication system that allows a number of independent

devices to communicate with each other." A WAN is one that covers a

much larger area (e.g., nationwide or worldwide) , and has one or more

computer nodes that are central to the operation of the network.

These specialized computer nodes support the routing -- storing and

forwarding -- of packets of information.

The simplicity of Local Area Networks makes them useful for

specialized applications within a small organization. They can

continue to operate with some of their devices broken or down because

any one unit does not affect the operational status of others.

Moreover, LANs promote and extend a cooperative work environment for

both people and machines.

When discussing LANS, an understanding of the following terms is

important:

Centralized main or host computer that does all

data processing;

Distributed     some remote computers do their own  
processing;

Gateway        hardware and software for two  
technologically different networks  
to communicate with each other;

Bridge         linking two technically similar  
networks to one another;

Servers        network peripherals that support  
specialized use by the entire  
network community, e.g., file  
storage servers and printers.

These elements make up a multilayered communications facility  
that represents a multitude of telecommunications networks that must  
interoperate on both a national and a global scale. Because  
telecommunications have developed in different ways in various foreign

countries, there has been a continuing pressure for standards and for the cooperation of all countries in the efforts of the (CCITT). This overall international telecommunications

57

environment supports a communications arrangement that may be logically segmented into:

1. A public communications network layer. The public network (at least, in the United States) is required to provide uniform service of good quality and on an equal access basis. It must permit uniform management of the network across the nation, and it must exhibit acceptable reliability characteristics to the public user. The regional Bell operating companies provide local public telephone service.

2. A business communications network layer. In this category, the communications structure is privately owned and operated. There are a multiplicity of these proprietary networks developed by private companies to reduce the communications costs to corporations. Tymnet, AT&T; and the regional holding companies assist corporations in building such private structures. Private networking will most likely increase in the future, but it may be implemented as virtual private circuits using the intelligent digital networks (IDNS) of the 1990's.

3. A business distribution network layer. This type of network transmits from one site and is received by many sites. Cable TV and the broadcasting of commercial television shows are examples.

The evolution of the IDNS must support the following operational characteristics of the local and national telecommunications system:

1. The current arrangement of public telephone networks and

packet switched networks does not support the simultaneous operation of voice and data services. The simultaneous transmission of speech, data, telemetry and signalling will be natural in future IDN networks.

2. The message content must be transparent to the various- services employed by the network.
3. The embedded base of existing network equipment must be accessible- by the evolving IDN. Such things as classical two-wire telephony must be supported.
4. The security and privacy of information must be available for all users of the network.
5. The appropriate levels of network management for handling

accounting, performance, configuration control, reliability and security of information must be available on the network.

## Planned Systems

The ultimate evolution of the current intelligent digital network is the Integrated Services Digital Network (ISDN) which has been emerging in the industrialized nations for the last ten years. It is a technology that ultimately will place end-to-end digital signalling capability throughout the network. It has been slowed because of two major factors:

1. The lack of standardization between vendors of transmission equipment within the United States and Canada, as well as widely divergent option selections that are specified by CCITT in terms of its so-called ISDN standard reference model. This latter situation has resulted in the inability of the Postal, Telephone and Telegraph Agencies of various nations including the United States to establish ISDN environments that could exchange information. The ability

to exchange information is called interoperability. Two ISDN networks that can exchange information transparent to two end users, one on each network, are said to be able to interwork.

2. The enormous established base of analog switching equipment.

This base is measured in the 10's of billions of American dollars and represents an investment by service providers such as AT&T; and end user organizations that cannot simply be replaced in a short period.

The United States government through the Brooks Act of 1987 has mandated that all agencies of the government must move to a common communication backbone that is to be an ISDN environment as soon as acceptable standards can be put in place. The National Institute of Standards and Technology (NIST) has been actively pursuing the realization of standards since February, 1988. The General Services Administration (GSA) with the awarding of the FTS2000 contract to AT&T; and Sprint is now working to develop an ISDN migration-plan that will be acceptable to all government agencies. This plan may have to



proceed on an agency-by-agency basis because different agencies will have unique problems in their telecommunications environment. The result is to be an intelligent network that will offer many services using digital signalling, and that will provide individual users with an extremely friendly

59

interface with their ISDN workstations (i.e., handsets, PCI's, integrated voice, data, and video consoles).

To the user, the ISDN environment appears as a highly intelligent network in which, aside from the network access points, no clear distinction can be made as to where their personal computer or mainframe ends and the network begins - in a sense, the computer becomes a part of the network and the network appears as a geographically dispersed computing environment. In essence, the intelligence that resides in the individual switching machines is made

available to the users of the network as a menu of services which can enhance the capability of the user to do a variety of functions. In an attempt to capture the needs of the user, NIST and the industrial telecommunications community created the North American ISDN User Forum in the Spring of 1988. This forum has been generating user applications for ISDN. As of June 1989, 81 applications had been cataloged.

Because of the high-level of intelligence invested in the ISDN environment, such concerns as user authentication at both the sending and receiving ends, end-to-end integrity of a message, and security of the information sent, can be dealt with by the network in a manner transparent to the user.

It must be recognized that the ISDN environment is a multimedia services facility that allows end-to-end transport of voice, data or slow-scan video. Facsimile (FAX) transmission is also part of this media mix. The current ISDN implementations in North America can support a maximum bit rate per channel of 64 Kbits/sec. This is called narrowband-ISDN. A separate standardization process is also

taking place in North America and around the world. It is called broadband-ISDN with anticipated bit rates more than of 600 Mbits/sec, an increase by a factor of roughly 10,000 over narrowband-ISDN. This network will provide services with an ultimate impact on the business and commercial customers of North America that will be larger than all the capabilities now associated with narrowband-ISDN. The usage of broadband-ISDN, in conjunction with rewiring the North American continent with fiberoptic circuits, will revolutionize information processing.

With the emergence of a single, seamless ISDN communications fabric the proliferation of private networks should be greatly reduced in both private industry and the Federal Government. This should substantially reduce the costs, of network operations, administration and maintenance. in particular, one governmental agency has estimated an annual cost savings of \$7 million in lust moving to an ISDN environment in terms of the reduction of network management charges. These savings do not address the potential increases in productivity through the acquisition of the new user services provided by an ISDN facility. The NIU-Forum is considering the cost-benefit concerns of organizations they move to a fully-ISDN equipped telecommunications

environment. This work

60

helps the unsophisticated user to use the intelligent network to carry out well-defined functions such as efficient data collection.

A further aspect of an ISDN environment is that the network could act as a highly intelligent protocol converter. In a sense, it could function as a concurrent multiple gateway between many different types of data networks. Uploading and downloading of data would be taken care of automatically and in a manner transparent to the users.

Verification of the data sent on an end-to-end basis also would be done automatically by the network. In a multi-media environment media conversions (voice-to-data, data-to-image, image-to-data, data-to-voice, and image-to-voice) also could be done by the ISDN facility.

The key here is the high intelligence of network, and the transparency of the ISDN operations to its attached user community.

V. REFERENCES

A. CATI

Curry, Joseph; "computer Assisted Telephone Interviewing: Technology and Organization management"; Sawtooth Software; June 17, 1987.

Groves, Robert M.; editor et al; Telephone Survey Methodology; John Wiley & Sons; 1988.

Nicholls, William L.; "The Impact of High Technology on Data Collection"; CATI Research Report No. GEN-1; Bureau of the Census; February 24, 1989.

Werking, George; Tupek, Alan; and Clayton, Richard; "CATI and

Touchtone Self-Response Applications for Establishment Surveys";

Journal of official Statistics; Vol 4; No. 4; 1988; pp 349-362.

#### B. CAPI

Danielsson, L.; and Maarstad, P.A.; "Statistical Data Collection with

Handheld Computers - A Test in Computer Price Index"; Unpublished

report of Statistics Sweden; Orebro, Sweden; 1982.

National Center for Health Statistics; "Report of the 1987 Automated

National Health Interview Survey Feasibility study - An Investigation

of CAPI": November, 1988.

National Center for Health Statistics and Bureau of Census; "Report of

the 1987 Automated National Health Interview Survey Feasibility Study,

An Investigation of Computer Assisted Personal Interviewing"; U.S.

Department of Health and Human Services; National Center for Health

Statistics; November, 1988.

Netherlands Central Bureau of Statistics; "Automation in Survey

Processing"; Select Report 4; Central Bureau of Statistics; Voorburg,  
Netherlands; 1987.

Nicholls, William L.; "The Impact of High Technology on Data  
Collection"; CATI Research Report Number GEN-1; U.S. Department of  
Commerce; Bureau of Census; , February 24, 1989.

Rice Jr., Stewart C.; Wright, Robert A.; and Rowe, Ben; "Development  
of. Computer Assisted, Personal Interview for the National Health  
Interview Survey 1987"; Proceeding of the Survey Research Methods  
Section, American Statistical Association; 1988.

Rothchild, Beth B.; and Wilson, Lucy B.; "Nationwide Food consumption  
survey 1987: A Landmark Personal Interview Survey

Using Laptop Computers"; Proceedings of the Bureau of the Census  
Fourth Annual Research Conference; pp 347-356; U.S. Department of  
Commerce; Bureau of the Census; 1988.

Sebestik, Jutta; Zelon, Harvey; DeWitt, Dale; O'Reilly, James M.; and McGowan, Kevin; "Initial Experiences with CAPI"; Proceedings of the Bureau of the Census Fourth Annual Research; pp 357365; U.S. Department of Commerce; Bureau of Census; 1988.

van Bastelaer, Alois; Kessemakers, Frans; and Sikkel, Dirk; "Data collection with Hand-Held Computers: Contributions to Questionnaire Design"; Journal of official Statistics; Vol.4; No. 2; pp 141-154; 1988.

C. CASI

Clayton, Richard, L.; and Winter, Debbie L.S.; Voice Recognition and Voice Response Applications for Data Collection in a Federal/State Establishment Survey"; Official Proceedings of Military and Government Speech Tech '89, Media Dimensions; November, 1989

Ponikowski, Chester; and Meily, Sue; - "Use of Touchtone Recognition Technology in Establishment Survey Data Collection"; Presented at the



First Annual Field Technologies Conference, St. Petersburg, Florida;

1988.

Werking, George; Tupek, Alan; and Clayton, Richard; "CATI and

Touchtone Self-Response Applications for Establishment Surveys";

Journal of Official Statistics; Vol 4; No. 4; 1988; pp 349-362.

D. H -machine interfaces

Card, S.K.; Moran, T. P.; and Newell, A.; The Psychology Computer

Lawrence Erlbaum Associates; Hillsdale, NJ; 1983.

Conklin, Jeff ; "Hypertext: An Introduction and Survey"; IEEE

Computer; pp 17-41; Sept, 1987.

Draper, Norman D.; User Centered System Design; Lawrence Erlbaum

Associates; Hillsdale, NJ; 1986.

Hartson, H.R. (ed); Advances in Human-Computer Interaction; Ablex

Publishing Co.; Norwood, NJ; 1985.

Myers, Brad A.; *Creating User Interfaces by Demonstration*; Academic Press;

San Diego, CA; 1988.

Shneiderman, Ben; *Designing the User Interface*; Addison-Wesley;

Reading, MA; 1987.

Shu, Nam; *Visual Programming*; Van Nostrand; New York, NY; 1988.

#### E. Computer Security

Department of Defense; *Trusted Computer System Evaluation Criteria*;

DoD 5200.28-STD; 1985.

Federal Information Processing Standards Publication (FIPS PUB) 39;

*Glossary for Computer Systems Security*; February, 1976.

Federal Information Processing Standards Publication (FIPS PUB) 461;

Data Encryption Standard; January, 1988.

Federal Information Processing Standards Publication (FIPS PUB) 73;

Guidelines for Security of Computer Applications; June, 1980.

Federal Information Processing Standards Publication (FIPS PUB) 112;

Standard on Password Usage; May, 1985.

Federal Information Processing Standards Publication (FIPS PUB) 113;

Standard on Computer Data Authentication; may, 1985.

Gasser, Morrie; Building a Secure-Computer System; van Nostrand

Reinhold; New York; 1988.

National institute of Standards and Technology Publication List 91;

Computer Security Publications; January, 1988.

Pfleeger, Charles P.; Security in Computing; Prentice Hall; New

Jersey; 1989.

P. Networks

Arni, D.; "Standards in Process: Foundations and Profiles of ISDN and OSI Studies"; National Telecommunications and Information Administration; Report 84-170; U.S. Department of Commerce; Washington, DC; December, 1984.

Browne, T.; "Network of the Future"; Proceedings of the IEEE; September, 1986.

Lutchford, J.; "CCITT Recommendations on the ISDN: A Review"; IEEE Journal on Selected Areas in Communications; May, 1986.

Madron, Thomas W.; Local Area Networks: The Second Generation; John Wiley and Sons; 1988.

Stallings, W.; Handbook of Computer-Communications Standards, Volume  
1: The Open System Interconnection (OSI) Model and OSI-Related  
Standards; MacMillan; New York; 1987.

Stallings, W.; ISDN: An Introduction; MacMillan; New York; 1989. U.S.  
Department of Commerce; "NTIA TELECOM 2000: Charting the Course for a  
New Century"; National Telecommunications and Information  
Administration; NTIA Special Publication 89-21; U.S. Department of  
Commerce, Washington, DC; October, 1988.

#### G. Applications

Clayton, Richard L.; and Harrell, Louis J., Jr.; "Developing a cost  
Model for Alternative Data Collection, Methods: Mail, CATI, and TDE";  
ASA Proceedings of the Section of Survey Research Methods., 1989.

Energy Information Administration; "PEDRO - Respondent User Guide to  
the Petroleum Electronic Data Reporting Option"; Version 3.0; February  
3, 1989.

Groves, Robert M; Survey Errors and Survey costs; John Wiley and Sons,  
New York, 1989.

Statistical Policy Working Paper 15; "Quality in Establishment  
Surveys"; Office of Management and Budget; July, 1988.

H. Standards

National institute of Standards and Technology Publication List 58;  
Federal information Processing Standards publications; June, 1989.

VI. Appendices

Appendix VI.A. Costs

Introduction

The choice of a collection method is usually based on a combination of performance and cost factors. For traditional methods, these factors are easily identified and the selection of a collection method is not difficult. With recent technological advances, new methods described in this report expand the array of potential collection tools and challenge the survey designer to reevaluate old cost and performance assumptions. The decision of which method or methods to use is now more difficult.

This section reviews the structure of costs in the data collection function covering several collection methods including mail, CATI, CAPI, TDE and VRE. It also briefly describes the impact of automated collection on costs, particularly versus mail operations. This profile of costs is limited to data collection; considerations of impact on sample design, questionnaire-changes, edits, and other issues are excluded.

#### Collection Methods Defined

CATI: The application of CATI is usually considered to address

timeliness and other quality problems. The computer assists by automatically controlling questionnaire branching, conducting on-line editing for reconciliation directly with the respondent, scheduling future calls and capturing a variety of management information about the interview. Thus, most data collection activities are conducted through the CATI system. The use of CATI generally vastly reduces or eliminates routine mail handling activities and postage costs. CATI adds new costs in equipment purchase and replacement and telephone charges.

CAPI: This method extends the benefits of controlled branching and on-line edit reconciliation to improve the quality of data collected by personal interviewing. In surveys already using personal visit collection, CAPI adds direct costs of computer hardware for each data collector and software design and maintenance.

Self-response -- Prepared Data Entry: By offering Prepared Data Entry to respondents, the collecting agency adds the costs of software design and maintenance, and possibly the costs of telephone charges for electronic transmission of the completed questionnaire.



Self-response -- Touchtone Data Entry and voice Recognition

Entry: These methods include many of the same sample monitoring

67

features of CATI and eliminate many of the labor-intensive activities associated with the traditional mail methods. TDE and VRE methods are currently used as a replacement of mail collection. By comparison, the regular mail handling to and from respondents is reduced to a single postcard to remind the respondent that it is time to call in their data. TDE and VRE further reduce manual operations by transferring key entry to the respondent. Short nonresponse calls may be employed to remind respondents to call in their data as publication deadlines approach. While reducing labor costs, TDE and VRE involve added costs for computer hardware and software development and maintenance.

## Cost Model

The data collection function is the series of activities that follow sample selection and precede estimation. Data collection is comprised of a series of activities for capturing the data, converting the data to machine-readable form, performing editing and edit reconciliation, and follow-up for nonresponse. The conduct of these activities varies greatly under mail, CATI, CAPI and self-response modes (PDE, TDE and VRE). Major recurring cost categories for these collection modes are outlined in Table 1.

Table 1. major Recurring Cost Categories by Collection Mode

Major Cost Categories	Mail	CATI	CAPI	PDE	TDE	VRE
LABOR						
mail out	x			x	x	x
mail return	x			x		
data entry	x	x	x			

edit reconciliation	x	x	x		x	x
nonresponse follow-up	x		x	x	x	x
software development/maint.		x	x	x	x	x
interviewer training		x	x	x		
NON-LABOR						
postage	x			x	x	x
telephones		x			x	x
computer hardware		x	x		x	x
travel			x			

The cost categories presented in Table 1 can be used to evaluate the costs of other collection methods. By comparing the activities of the alternative method to the current method, a rough determination of affordability can be made. Detailed cost studies would be necessary for each specific survey application.

## Assumptions

Realistic assumptions are a vital part of an analysis of costs.

Several assumptions should generally be made about the level of workload and equipment requirements. These may include the number of units per CATI interviewer during normal collection period, and the number of minutes per interview. The TDE cost assumptions include the length of the average call, effects of peak calling periods, the number of incoming lines per TDE board, and the average proportion of units receiving nonresponse prompting actions. Also, the number of boards that can be placed in the microcomputer should be included.

The following factors, independent of collection mode, should be included in the model: salaries and benefits, administrative overhead allocations, standard non-personnel services, postage, amortization of computer hardware to cover replacement, and telephone charges, including fixed monthly line charges and variable call costs.

The following factors are generally difficult to quantify and

often cannot be treated equally for all methods: start up costs for research and development, ongoing systems design and maintenance, training, and emergency back-up features for CATI and TDE.

#### Other Important Considerations

Critical decisions concerning changes in the data collection methods are not made solely on costs; there are many other considerations to include in these decisions.

Organizational Impact: The design of an effective production environment is essential to timely, ongoing output of data. For example, the success of CATI and TDE in compressing the collection period may pose peak period staffing problems. Also, the cost model assumes the managers can perfectly capture and reallocate resources as collection methods change. For example, TDE eliminates key entry. The costs are only truly saved if these resources can be captured and reinvested in new equipment and telephone charges, and with remaining savings redirected toward improving the quality of other survey

functions. Also, it is assumed that postage savings also are identifiable and may be similarly captured and redirected.

Staffing for Research and Development: The development of new techniques usually requires a small staff dedicated to achieving the change desired. Also, this staff must have a variety of skills, including economics, statistics, methods test design, computer systems design, questionnaire development, and analytical, writing, and presentation skills. This combination of individuals may be difficult to identify and remove from ongoing production

tasks. Given the frequency of new issues and problems, this group may require special attention from management and latitude in trying creative approaches to solving the wide range of problems that will inevitably arise in development efforts.

Systems Design, Programming, and Maintenance: There are significant start up costs, although these can be easily amortized over large, recurring surveys. These costs will vary with the complexity of the application and the experience of the development staff. Ongoing maintenance depends on the frequency and magnitude of the changes.

Training: Training requirements for staff to maintain manual operations, such as would be needed under mail, are small. Under CATI, a broader range of skills is required, including telephone communications skills and some working knowledge of the computer. The TDE system requires little special knowledge, keeping costs low.

Emergency Procedures: As we increasingly rely on technology to do work for us, we are increasingly at risk when it fails. All implementation approaches should include back-up procedures and equipment at appropriate locations to ensure uninterrupted service to respondents. Telephone based methods may require back-up computers and associated equipment standing ready for instant replacement. In addition, TDE and VRE applications consider establishing "call forwarding" services ready to route incoming TDE and VRE calls to an

alternative collection site if the primary collection computers malfunction.

### Quality Costs

The costs of quality are notoriously difficult to identify.

Often, it is easier to invert this idea to address costs of poor quality. For example, address refinement workload for solicitation is a cost of poor quality in the sample frame. Some edit reconciliation activities compensate for poor quality of collected data that may stem from deficiencies in concept or questionnaire design. Efforts expended to prevent future costs of poor quality, while often difficult to justify, generally pay off in lower ongoing costs.

### Future Costs

The choice of collection mode, or which combination, will depend on the particular survey application and the existing cost structure. However, it is important to view investments in data collection over



the long-term as the relative costs of each of the above inputs do not remain constant over time. Table 2 shows recent annual data on cost trends for the major cost inputs.

70

Labor and labor-intensive inputs, such as postage, are increasingly more expensive, while capital-intensive factors, such as telephones and computers, become less expensive. Based on these data, and other historical cost trends, there may be a growing advantage to switching to collection methods that use less labor and more capital.

Table 2. Recent Annual Changes in Costs of Inputs into Data Collection

Cost Category    Recent Annual Cost Changes (source)

Labor:            +5.8% for state and local government employee

                  compensation (ECI for the 12 month period ending June

1989)

Postage: +4.5% for the 1st class postage (U.S.P.s. for the rate  
increase in April 1988 to 25 cents)

Telephones: -1.3% for interstate toll calls (CPI-U unadjusted  
change December 1988 to December 1989)

-2.5% for intrastate toll calls (CPI-U unadjusted  
change December 1988 to December 1989)

Travel: +3.9% for private transportation (CPI-U unadjusted  
change December 1988 to December 1989)

Computers: -10.0% for microcomputers (PPI experimental price  
indexes for the 12 months ending January 1990)

Survey managers should project unit costs for their surveys for  
alternative collection methods over a ten year period using recent  
price trends. This approach illustrates that decisions to implement  
alternative methods should be viewed in terms of estimates of future

price levels. Decisions on conducting research and development testing need not await a current favorable cost benefit situation.

## Conclusion

The decision on exactly how to use each collection mode will vary by survey application. For example, CATI and TDE could be combined to address chronically late mail respondents. These units will first be converted to CATI collection to improve their reporting behavior in terms of timeliness and accuracy. These units will remain under CATI collection for about 6 months; a period adequate for reducing nonresponse problems, determining exact data availability dates (for subsequent nonresponse prompting), educating respondents to the importance of their data

and reinforcing timely reporting behavior. Then, the units will be converted to TDE collection to reduce costs while retaining sample

control. voice recognition collection could be used for those units without touchtone phones or for those respondents who prefer voice collection.

The approach outlined here is a basic tool for survey managers in assessing the potential application of new collection methods. Survey researchers should not be dissuaded by current costs from considering the use of automated collection methods. Recent cost trends suggest that the cost-effectiveness of collection methods changes over time. This should be considered in decisions concerning choice of collection methods for the future.

#### Appendix VI.B. Quality Improvements offered by CASIC

Quality problems generally result from inadequate planning or control of one or more steps in the survey process. CASIC cannot

replace or compensate for poor planning, but it may offer vast improvements in control by reducing manual intervention, promoting consistent procedures, by using supplementary data sources, and on-line editing to improve the accuracy of the data collection process.

The automation of the questionnaire is the primary way casic improves control, by offering consistent procedures, on-line editing, and use of other information, to monitor and control the interview which otherwise would have proven too difficult or burdensome on the interviewer.

While CASIC offers the potential for improvements, actual reductions in error components can only be made through efforts to delineate error potential and incorporating specific error-reducing techniques in the questionnaire.

Some error reductions may be great, and others may be small. However, none will result without thorough evaluations of error sources and planning to address each. Often, knowledge of the magnitude of various errors may be necessary to decide on the cost-

effectiveness of addressing some error sources.

The automation of the data collection process directly reduces some sources of error. For example, telephone collection of data may reduce the potential for processing error resulting from mailing the wrong form to a respondent. Other indirect benefits can be obtained through automation, including reductions in coverage error. For example, on-line evaluation of respondent characteristics provides immediate identification of out-of scope respondents.

This section discusses several error components that may be reduced through CASIC methods. The structure, definitions and background of this discussion were derived from Statistical Policy Working Paper 15, entitled "Quality in Establishment Surveys." Readers are encouraged to refer to this document for more information on error definition, sources, control methods and measurement aspects.

#### Specification Error

Specification error occurs at the planning stage of a survey when

specification is inadequate or inconsistent with the objectives of the survey. It can result from the difficulty of measuring abstract concepts or from poorly worded questionnaires and instructions.

73

CASIC methods may reduce specification errors in several ways.

For example, difficult concepts may require very detailed questionnaires with complex branching patterns to obtain correct measures. CATI and CAPI can allow greater flexibility in structuring questionnaires than would be possible using paper forms. Also, CASIC provides a means for correcting specification error once identified.

If one or more questions are difficult to use during collection, or responses seem improper, corrections can be made centrally and software transferred quickly to all collection points. Given printing timing and costs, use of paper forms probably would not allow such mid-stream changes and the survey results could be severely compromised.

While developing and printing questionnaires, skip pattern indicators may be omitted, or the tedious work of proofreading multiple variations may lead to errors. Use of CASIC instruments are just as susceptible to this error as are written forms. Automated questionnaires, and the associated code, must be checked thoroughly to ensure their accuracy. Forms also may be faint or smudged leading to difficulty for the respondent.

Traditional methods for measuring specification error include record check studies, cognitive studies, questionnaire pretests, and comparison of results with independent estimates. CASIC can contribute to these approaches. First, record check surveys that scrutinize detailed definitional areas may be very complex. Such detailed branching is a strength of CATI and CAPI.

Coverage error

Coverage error includes both undercoverage, the exclusion of in-scope units; and overcoverage, the inclusion of out-of-scope units.



CASIC may reduce overcoverage if the questionnaire includes checks for scope-determining characteristics. Data for sample units failing these criteria may be noted for review or exclusion, or the interviews may be ended rather than waste time. Also, duplication errors, stemming from duplicates on the sample frame may be identified through an automated records review at any point during collection. Again, such benefits are only possible with initial planning.

#### Response Error

Response error is the difference between the correct value and the value collected. Respondent error is the failure to report the correct value, and interviewer error is the failure to record the data properly.

Respondent error may be controlled by comparing current data to previously reported data. Such on-line logic and internal

consistency edits can identify and resolve response errors directly with the respondent, rather than waiting for post-collection editing to catch errors for often spotty reconciliation follow-up. The power of an automated questionnaire also reduces interviewer error through instantaneous editing on any data entry mistakes large enough to trigger edit failures.

Interviewer consistency also may be controlled by monitoring interviewer practices and assuring conformance with specified procedures. Most large, centralized CATI facilities allow supervisors to listen to interviews in process and to view screens simultaneously.

#### Nonresponse error

Nonresponse errors follow from failures to collect complete information from all units in the selected sample. There are three types of nonresponse error: noncontacts, unit nonresponse and item

nonresponse. Each can be addressed through CASIC methods.

Noncontacts of selected units may be the result of interviewer

oversight, failure to locate the designated respondent due to

incorrect address or telephone number, or failure to get the form to

the respondent. CASIC cannot address weaknesses in mailing procedures

except by replacing them with accurate telephone contact. This would,

of course place additional burden on the accuracy of telephone

numbers.

interviewer oversight would be addressed by monitoring sample

status data that can be collected during-interviews. For example, a

detailed CATI system may capture information each time a call is

placed, the number of attempts made to each number and the result of

each attempt, such as "no answer" or "busy." Noncontacts may then be

classified as not attempted versus unsuccessful attempts.

Unit nonresponse occurs when no information is received from the

respondent. The survey designer must strive to make reporting as easy

as possible to reduce intentional nonresponse. Almost any effort that

improves the respondent's understanding of the survey is worth the

cost. The convenience of reporting is essential, as is the clearest

and shortest possible interview. One CATI application reduced sample attrition by over one third compared to mail, attributed mostly to strong scheduling and building strong rapport with the respondent and providing information about the importance of the survey and its timing needs.

Item nonresponse occurs when the respondent does not answer certain questions during the interview. This error may occur when the respondent cost of compiling data is too great, or the data are not easily available during the collection period. of course, some data may be sensitive or confidential.

Item nonresponse also may occur through the failure of the interviewer to ask questions or follow procedures. It is in this area that CASIC is most beneficial. By using software to control interviews, CATI and CAPI interviewers are not allowed to make errors

of omission or purposely to skip questions.

Another important part of reducing item nonresponse is to use a priori knowledge about the respondent. For example, in establishment surveys, information about the record keeping practices of the respondent may be retained on the computer for access during the interview that could provide special branching to elicit firm-specific data. This approach would generally be too cumbersome without computer assistance.

#### Processing Error

Processing error stems from the faulty use of correctly designed survey methods. It encompasses many collection and post-collection errors and the printing of the questionnaires. Also, processing error may arise from clerical handling of forms whether in mailing or key entry.

CASIC methods, by reducing or eliminating these labor intensive

and error-prone activities, can substantially reduce processing errors. CASIC respondents in recurring surveys may receive a mailed form once per year rather than once each month or quarter, reducing the opportunity for mail-related errors. All CASIC methods ensure that data entry and other coding is done by a well trained interviewer or by the actual respondent, thus reducing keypunch error. All CASIC procedures should include repetition of the incoming data for verification with the respondent. CATI and CAPI interviewers repeat the data aloud as they are keying it, and CASI methods must provide for repeating the data for verification. by the respondent. on-line edits again play a role in. assuring that data errors are caught before they get to the post-collection stage.

Another source of processing error is data processing by computer. All the benefits of CASIC methods described above may be diminished by errors in computer processing. Failures in designing and constructing CASIC methods may substantially reduce data quality. For example, poor branching or non-exhaustive response options may prevent knowledgeable interviewers or self-response systems users from properly completing interviews.

Quality, as discussed above, is often defined in terms of statistical error or lack of accuracy. However, the idea of quality contains several other elements. For example, the element of timeliness is critical to most surveys. Accurate data that are too late to be of use have little quality. The use of CASIC methods, like CATI and TDE, has proven useful in improving the

76

timeliness of data in one large establishment survey, thus offering the potential to reduce the number and magnitude of estimate revisions. Quality also includes costs. Two identical products with differing costs are of different quality.

Also, quality control should be applied to the process of methods development. A high quality CASIC application must be easily understood and easily used by interviewers and respondents. Anything less is of low quality.

## Conclusion

CASIC methods have great potential for improving the control over data collection activities and the quality of the resulting data as it moves toward the post-collection survey functions. This discussion of survey error and the application of CASIC methods is not exhaustive of either current or potential approaches. Many other creative approaches will be developed to further use the power of computers to aid in improving the quality of Federal surveys.

Equally important to add to the discussion of quality is a caution that the mere use of CASIC methods does not automatically guarantee higher data quality. Failures in designing and testing questionnaires or in using other standard survey practices will inevitably result in data quality problems.

The increased reliance on software development has important implications for hiring and training skilled survey designers. Statistical methods knowledge and experience alone are not sufficient



qualifications to achieve satisfactory results. Previously distinct boundaries between occupational groups will continuously blur or disappear. In the future survey design will likely be increasingly accomplished through teams of skilled workers from different occupations. Just as statisticians must be familiar with software design techniques to understand their implications, systems analysts and programmers must be familiar with the statistical aspects of the survey and questionnaire design. Managers of automated surveys cannot avoid having a background in all aspects of the design, implementation and maintenance of integrated systems.

#### Appendix VI.C. Survey Examples

The following examples provide additional examples of current CASIC applications. Each provides a point of contact for additional information.

National Agricultural Statistics Service (NASS)

Agricultural Surveys

Collection Type -- CATI

Point of Contact

USDA - NASS

CATI Section, Survey Management Branch

Research and Application Division

1400 Independence Avenue

Washington, DC 20250

Type of Data to be Collected

The Agricultural Surveys are conducted in January, March, June, July, September, and December to collect data on crops, livestock, grain stocks, and other information from farmers. Starting with the March 1987 survey, data were collected using Computer Assisted Telephone Interviewing (CATI) to replace the paper-and-pencil mode. CATI is a computer driven telephone interviewing system developed to replace a paper questionnaire with a more efficient, error-reducing questionnaire. It can edit the data as it is entered by accepting only valid responses; checking sums and edit limits; carrying forward responses required for subsequent questions; and refusing answers inconsistent with current or historical responses. CATI provides question branching and some systems can handle each state's customized version of the Questionnaire. Currently, 14 of the 45 field offices are collecting data with CATI using 183 calling stations, and in 1989, over 70,000 farmers were contacted to obtain Agriculture survey data. CATI usage will expand rapidly with the installation of the new PC Local Area Networks (LANs) in the-field offices. By 1992, all 45 field officer. will be equipped with a PC LAN and there should be about 750 calling stations available -for making CATI calls.

## Approach to Respondents

The Agriculture Surveys CATI application is written using the Computer Assisted Survey Execution System (CASES) software developed by the University of California at Berkeley. CASES has an automated sample delivery system that is in use and an automated call scheduling and dialing option will be initiated in the future. Other features make CASES one of the most powerful systems on the market today. These include: interactive editing (coding), sample management, records keeping, conversational survey Analysis (CSA), audit trails, jump-back menus, and full screen mode with cursor control. The interview sessions are initiated by the interviewer.

The computer program controls branching to or skipping among questions, and validates the data as it is entered. In addition, the interviews are more personalized, probing questions are standardized,

use of historic data is standardized, and the questions can be more sophisticated than those on paper questionnaires.

#### Transmission

Data collected via CATI is currently up-loaded to an IBM mainframe leased from the Martin Marietta Corporation where a SAS edit is done, and data summarized. Since the survey data is currently collected via different modes (CATI, telephone, on paper, personal interview, and mail), it is necessary to convert the data to one standard system for summarization.

#### Factors Affecting Choice of Method

The implementation of the CATI for collecting Agricultural Survey data has resulted in higher quality data and a reduction in time and cost of collection. This is due to combining the collection, entry, validation, analysis, and conversion of data. More complex questionnaire design is possible since the program controls branching

and logic. CATI works particularly well in situations where a short implementation schedule exists.

#### Quality Issues

Significantly fewer errors occur, as data is validated at the time it is reported and keyed. The data validation currently includes internal data checks but some work has been done on using historic edit checks as well. Since the program controls the logic, you are assured that all questions are asked consistently. A totally menu driven system is being designed and will be in operation soon.

National Health Interview Survey (NHIS)

Computer Assisted Personal Interview (CAPI) Case Study

Collection Type -- CAPI

Point of Contact

Division of Health Interview Survey

National Center for Health Statistics

3700 East-West Highway

Hyattsville, MD 20782

(301) 436-7085

Type of Data to be Collected

The case study involved the collection of health data from approximately 500 households in two Census Regions: Chicago and Charlotte. The questionnaire consisted of the NHIS core questionnaire that contains more than 600 questions on the composition of the household, demographic characteristics, health status of the individuals, health care visits and incidents, and other pertinent health care data. The respondents are contacted at their residence, and are not contacted again unless the interview was not completed on the initial visit or additional clarifications are needed. Because

this effort was a feasibility study for CAPI, only a small portion of the normal survey respondents were contacted. The normal survey size is 50,000 households per year.

#### Approach to Respondents

CAPI was used to obtain the survey information. A portable computer containing the survey questionnaire was carried-into the household by the interviewer. The portable computer was a Toshiba 1100+ weighing approximately 10 lbs. The survey questionnaire was programmed in the Computer Aided Survey System (CASS) language developed by Dawn and Charles Palit at the University of Wisconsin. The interviewer conducted the survey by reading the questions from the computer screen and entering the answers on the keyboard.

#### Transmission

The survey questionnaire data is collected on 3 1/2" floppy disks by the interviewer. The disks are collected from each interviewer in the region, merged at the regional office, and then mailed to the



computer center in North Carolina for uploading to the mainframe computer.

#### Factors Affecting Choice of Method

The choice of CAPI provided several advantages. First, improved timeliness of survey data availability through the ability to quickly put the survey into the field and the subsequent elimination of the keying of the completed questionnaire.

Second, improved data quality because (1) significant editing can be done as a part of the data collection process; (2) there is greater flexibility for questionnaire design, e.g., more opportunity to make changes closer to the field implementation date; (3) good measurements for non-sampling error are easily provided as a part of the process; and (4) immediate interviewer quality control is available from an

analysis of the data, e.g., time to complete a section or the entire questionnaire.

82

Current Employment Statistics Survey

Bureau of Labor Statistics

Collection Type -- CATI, TDE, VRE

Point of Contact

Division of Monthly Industry Employment Statistics

U.S. Bureau of Labor Statistics

Room 2089 441 G Street, N.W.

Washington, D.C. 20212

202--523-1446

## Type of Data to be Collected

The Current Employment Statistics (CES) survey collects data from over 300,000 nonagricultural business establishments each month covering employment, hours and earnings.

The CES is voluntary and is conducted in a Federal-State cooperative system in which BLS provides the statistical standards and procedures for use in each state and the District of Columbia, Puerto Rico and the Virgin Islands. In this way, the resulting data can be aggregated to National totals, and are comparable among the states, which produce estimates at the state and metropolitan area levels.

The national data are first published after only two weeks of collection. Then, based on additional sample receipt, revised estimates are published after 3 more weeks of collection, followed by final estimates after a total of 8 weeks of collection. The short collection period poses the toughest problem for the CES survey.

## Approach to Respondents

Under mail collection, respondents return the form sometime after their data become available. Given the very short, two week collection period before the publication of preliminary estimates, any delay in completing the form, or returning it to the state has severe implications for response rates. Under CATI collection, respondents are called on a pre-arranged date, if possible, the same day -as the firm's data are available. The data are entered and edited during this call, and the next month's call is scheduled.

The conversion of respondents from mail to CATI includes sending selected units a package of materials with information on the importance and uses of the CES, data, and instructions on

reporting by telephone. As respondents are converted to TDE or VRE collection, another package is sent containing instructions on how to

participate using these methods.

Under TDE and VRE, respondents receive an "Advance Notice" postcard during the reference period that serves as a reminder that it is time to call in their data. The collection microcomputer is available 24 hours, 7 days a week to receive calls. A few days before the end of each collection period, the TDE and VRE collection files are checked, and those respondents for which data are missing receive a short call to ask that the data be called in.

After the first month of collection by TDE or VRE, respondents are called to discuss the new method, to identify and correct any problems that may have been encountered, and to insure trouble-free collection.

#### Transmission

Under the mixed mode of collection in the CES program, responses are received by mail, CATI collection and TDE self-response. In the

Federal/State cooperative system, the state collects the microdata, through the appropriate mix of methods, for electronic transmission to the central computing facility in Washington. The State data are then aggregated for the production of national estimates. At each level, the microdata are subjected to rigorous editing for logical, consistency, and longitudinal checks.

#### Factors Affecting Choice of Method

#### Timeliness

BLS has been conducting research and development in the area of computer assisted methodology since 1984. Currently, over 5300 units are collected via CATI each month.

The use of CATI within the CES program is limited by the resources available. The current implementation strategy is based on targeted use of CATI for specific segments of the sample which warrant special treatment and commitment of funds. These segments include large, "certainty" units, and late respondents. These units are

converted to CATI collection for a short period, usually 6 months, to educate respondents on the importance of the CES data and the reporting timing requirements and to improve reporting habits. After reporting improves, these units will be returned to either TDE self-response collection, or mail, if there is no access to a touchtone phone. Thus, CATI is seen as a transitional tool

84

for improving the overall timeliness of the CES sample over a period of just a few years.

#### Costs

While CATI is a very strong method for improving the timeliness, it is currently more expensive than the mail collection process that has been used for decades. The high costs of CATI prompted BLS to pursue development and testing of TDE and VRE methods. These automated self-response methods offer lower costs through reducing or

eliminated many manual activities and postage involved in mail collection. Respondents without touchtone phones will be collected using voice recognition.

### Quality Issues

By every measure, CATI proved superior to mail collection, and TDE has shown the ability to maintain high response rates over extended periods of more than two years. The tests of VRE collection show similar ability to maintain high response rates.

Performance Measure	Collection Method		
	Mail	CATI	TDE/VRE
Sample received for:			
preliminary estimates	50%	85%	85%



revised estimates	75%	99%	99%
final estimates	87%	100%	100%
Sample attrition			
(annual rate)	10-15%	2-4%	2-4%

Besides reducing nonresponse error for the preliminary estimates, the CES program uses a CATI system to evaluate and correct response error. Large scale tests using telephone record check surveys have shown that this approach is useful for insuring that the reported data conforms as closely as possible to CES definitions.

Collection Type -- PDE

Point of Contact

Reserves and Natural Gas Division

Energy Information Administration

1114 Commerce St., Room 804

Dallas, Texas 75242-2899

(214) 767-2200

Type of Data to be Collected

There are approximately 600 respondents who are oil or gas well operators who produce at least 400,000 barrels of crude oil or 2 billion cubic feet of gas annually. There are 15 detailed questions in this annual survey. A system of reporting on PC diskettes was set up on an operational test basis for the collection of 1988 data. Ten percent of 1988 production was reported with RIGS.

## Approach to Respondents

The questionnaire runs on IBM PC compatible computers with at least 360K bytes of RAM and two floppy drives or a floppy drive and a hard disk drive. The user only needs to know basic Dos functions. The program is menu driven, and on-line help is available as well as a toll-free telephone hotline during business hours. It comes with a fifty page Users Guide.

## Transmission

Respondents copy the data files onto a floppy disk and mail the disk (with the cover page sent to them) to EIA. They also have the option of sending in the original paper form.

## Factors Affecting Choice of method

RIGS was developed to provide respondents with an alternative, more user-friendly means for reporting data. The PC compatible computer was chosen because of its wide availability. Use of the mail avoids security concerns about data transmission. EIA processing is

done on a secure machine.

86

#### Quality Issues (Human Interface)

RIGS includes data edit checks to prevent inadvertent entries and an on-line correction capability. Company totals are automatically calculated. Respondents are requested to keep a copy of the data files and a printed copy of the output in case EIA's quality control analysts need to contact them. Reduction of follow-up calls is a significant benefit.

87

Electronic Filing System Office

Collection Type -- PDE

Point of Contact

Operations and Marketing Branch

Electronic Filing System Office

Internal Revenue Service

1111 Constitution Avenue, N.W.

Washington, DC 20224

(202) 535-6394

Type of Data to be Collected

In the early 1980's, the Internal Revenue Service (IRS) decided that the electronic transmission of returns by tax preparers to IRS would be both a practical and cost-beneficial alternative to the mailing of paper tax returns when a refund is claimed. According to

the Agency, the benefits of electronic filing would include: (1) reduced manual labor costs required to process, store, and retrieve returns, (2) faster processing and retrieval of tax data, and (3) reduced interest IRS is required to pay to taxpayers who file timely refund returns, but who are not issued refunds within the interest-free period allowed to the IRS to process these refunds.

Further, IRS reports show that electronically transmitted returns are processed with significantly fewer errors than paper returns. According to IRS figures for the 1988 filing season, as of April 29, 1988, 20 percent of paper returns processed by IRS had errors and only 5.5 percent of those filed electronically had errors. For taxpayers, electronic filing can mean refunds up to 3 weeks sooner, and because IRS can deposit these refunds directly into taxpayer bank accounts, refunds may arrive 3 to 4 days earlier than that. For tax preparers, the ability to provide electronic filing services to taxpayers promises a competitive business edge.

Approach to Respondents

in 1986, the program was initially tested in three metropolitan areas, and five preparers electronically filed 24,820 returns to the Cincinnati Service Center. In 1987, 69 preparers in 7 metropolitan areas electronically filed 77,612 returns. For the 1988 filing season, IRS expanded its electronic filing program to 16 IRS districts and a second service center in Ogden, Utah. With the expansion in 1988, the number of preparers increased to 2,339. Of that total, 1,114, or about half, filed all of the 583,077

88

electronic returns for 1988. Furthermore, H & R Block offices accounted for 82 percent of the total returns filed electronically during the 1988 filing season.

Transmission

To operate electronic filing at each of the two service centers

in 1988, IRS bought the International Business Machines Corporation (IBM) Series I computer, a local area network, and the related computer software. The network has IBM and IBM-compatible personal computers, high-resolution graphics display workstations, laser printer, tape drives, and optical disk drives. IRS uses the Series I to receive preparers' transmissions of electronic returns and to transmit certain information to preparers. The local area network was expected to do two primary functions: (1) retrieve and visually display the electronic returns on the tax examiners I workstations for error correction, and (2) permanently store these returns.

The basic components needed to prepare and transmit electronic returns include a computer, IRS-approved software to prepare tax returns, and the communications equipment and IRS-approved software to transmit the returns to IRS. In addition, IRS tests and verifies the preparers' competence in transmitting electronic returns.

The electronic filing process begins when a preparer transmits electronic returns to the service center. The Series I receives the transmission and writes the data onto a magnetic tape. The tape is



then manually transferred from the Series I to the service center mainframe computer processing. The mainframe generates an acknowledgment file specifying the received returns and whether each is accepted or rejected, and then writes this file onto magnetic tape. This tape file is hand carried from the mainframe to the Series I for electronic transmission to the individual preparers. Mainframe processing also identifies electronic returns containing errors - After IPS corrects the errors, tapes containing data from accepted error-free returns are sent with data from returns filed on paper to the IRS National Computer Center in Martinsburg, West Virginia, where the master files of tax account data are updated.

Energy Information Administration

Petroleum Electronic Data Reporting Option (PEDRO)

Collection Type -- PDE

Point of Contact

Petroleum Supply Division

Energy Information Administration

1000 Independence Avenue, S.W.

Washington, D.C. 20585

Type of Data to be Collected

The Petroleum Supply Division (PSD) of the Energy Information Administration (EIA) decided in 1987 to investigate electronic forms submission to collect the Petroleum Supply Reporting System (PSRS) survey forms. Ten of the major petroleum companies who file the mandatory "Monthly Refinery Report" were contacted to assess their PC and communications capabilities. The respondents contacted showed interest in investigating the use of PC's to collect this data. Most of these were already using PC's for business, personal, or academic purposes. The respondents either had a PC in their office area or had access to one in another office. Software such as Lotus 1-2-3 and

Dbase III could usually be found on these PC's. Some PC's were equipped with communications capabilities and those respondents were already using telephone lines for company reporting. It appeared to be the appropriate time for the PC to enter the PSRs data collection process.

#### Approach to Respondents

Early in 1988, PSD developed the Petroleum Electronic Data Reporting Option (PEDRO) and began providing its respondents with a software diskette by which they could create an electronic image of the form on a PC screen and enter their data in the appropriate cells. Firms having the necessary software capabilities can use their database to feed the data directly to the electronic survey form, eliminating keying and transcription errors. User-friendly software with help functions has been added to data entry functions to provide quick reference to definitions, conversion factors or other information to speed the completion of the survey form. This eliminates the need to search hard-copy files for survey forms instructions, product definitions, conversion tables, etc.

## Transmission

The data received on EIA survey forms are subjected to rigorous edit tests before they are accepted for inclusion in the EIA database.

These data are later summarized to produce EIA publications and reports used by the industry, the Congress, and the public.

Timeliness and accuracy are needed in every step of the data collection process. Collecting data via electronic means allows EIA to pursue another approach to saving time by providing respondents with electronic forms software that also does the survey edits and isolates anomalies for review before submitting the survey response to EIA. Issues which would require an EIA data analyst to contact a respondent by telephone for resolution are highlighted immediately.

This allows the respondent to correct any errors or attach a resolution indicator/comment to explain any anomalies. Additional telecommunications software has been added to allow the direct link

between the respondent's PC and the EIA system. Now the capability exists on a PC to create, quality check, and transmit an electronic file directly to EIA. This file is immediately accessed by EIA processing software and security and data transmission integrity tests are done.

The PEDRO software contains electronic forms for data entry, software for statistical editing and establishes a communications link between the respondent's PC and the EIA Computer Facility. The functions are menu-driven and use macro languages and script files to eliminate rudimentary tasks. The PEDRO system only requires that the respondent's PC run DOS software and be equipped with telecommunications capability.

Collection Type -- PDE

Point of Contact

Nuclear and Alternate Fuels Division

Energy Information Administration

1000 Independence Avenue, S.W.

Washington, D.C. 20585

(202) 254-5558

Type of Data to be Collected

The Nuclear and Alternate Fuels Division of the EIA conducts an annual survey of nuclear electric utilities that own commercial nuclear reactors. The BIA collects data on over 100,000 nuclear fuel assemblies that are owned and managed by these utilities. These data are collected in support of the programs of the Department of Energy's Office of Civilian Radioactive Waste Management. A system of reporting on PC diskettes was set up in 1986 and began with the

collection of 1985 data.

#### Approach to Respondents

The respondents are supplied with a program diskette containing compiled software and a data diskette. The data diskettes have the respondent's prior data submissions that are needed for comparison purposes and space for the current submission. The respondents load the program and data diskettes on their compatible PC's and enter the current data which is verified by the data entry program as it is keyed. They print a copy of the data submission, sign a certification statement for it, and return the printed copy and statement to the EIA with the diskette.

#### Transmission

The diskettes are mailed from the EIA to the respondents and the completed data diskettes are returned to the EIA by mail.

Telecommunication between the EIA and the respondents is not needed.

When the diskettes are received at the EIA, they are loaded onto a PC and checked. The data are uploaded from the PC's to the EIA mainframe

over local telephone lines. Note that since these data are for public utilities, they are in the public domain and thus not confidential or proprietary. Certain issues of data security do not apply for this survey.

92

The diskette form of submission is preferred, but not mandatory.

Respondents have the option of filing a paper form. Now there are approximately 70 utilities required to report for approximately 125 reactors, and all reports are filed on diskette.

#### Factors Affecting Choice of Method

The major advantages of the diskette collection are:

Data accuracy has been improved by (1). editing the data as they are keyed and (2) in some instances, data entry by technical



rather than clerical personnel. The second reason suggests that a higher level of technology in data collection may result in the availability of a higher level of respondent skill to complete the survey.

More data, including data of a more complex nature, can be collected using the diskettes compared to using paper forms.

Data are available sooner.

In planning such a system, government agencies must be careful to create a system that does not require or endorse a particular brand of hardware or software. Software licensing agreements also must be carefully reviewed to ensure they are not violated when software is provided to respondents.

## Appendix VI.D. A Taxonomy of Information Gathering Using a Computer

During this study there have been wide-ranging discussions on naming conventions for information gathering using a computer. The discussion has been so wide-ranging that the name of the committee has changed at least 3 times. This note was originally titled "Acronyms for Survey Technologies." However, it provides a good model of the different procedures for collecting information with computer assistance. The title of this section has been changed to reflect this model.

We can distinguish two aspects of the data collection process which may include automation: (1) assistance during the interview and (2) interaction with the respondent. A computer or other technology may be involved in one or both. Here is a system of acronyms using codes to show how each part is handled:

Operation types:

CA = computer assisted

MA = manually assisted

Interaction types:

PI = personal interviewing (person to person)

SI = self interviewing (respondent reads the questions)

TI = telephone interviewing (person to person on the phone)

TO = touchtone interviewing (respondent talks on the phone to a machine that discerns touchtones)

VI = voice recognition interviewing (respondent talks on the phone to a machine that discerns voices)

From these we get various possibilities, old and new:

CAPI = computer assisted personal interviewing

CASI = computer assisted self interviewing

CATI = computer assisted telephone interviewing

CATO = computer assisted touchtone interviewing

MAPI = manually assisted personal interviewing

MASI = manually assisted self interviewing

MATI = manually assisted telephone interviewing

A third aspect in some cases is how the data are sent to the processing center:

MA = mail

NE = network (wide area computer network)

TE = telephone line (direct line to computer)

The diskette form of submission is preferred, but not mandatory.

Respondents have the option of filing a paper form. Now, there are approximately 70 utilities required to report for approximately 125 reactors, and all reports are filed on diskette.

Factors Affecting Choice of Method

The major advantages of the diskette collection are:

Data accuracy has been improved by (1) editing the data as they are keyed and (2) in some instances, data entry by technical rather than clerical personnel. The second reason suggests that a higher level of technology in data collection may result in the availability of a higher level of respondent skill to complete the survey.

More data, including data of a more complex nature, can be collected using the diskettes compared to using paper forms.

Data are available sooner.

In planning such a system, government agencies must be careful to create a system that does not require or endorse a particular brand of hardware or software. Software licensing agreements also must be carefully reviewed to ensure they are not violated when software is provided to respondents.

## Appendix VI.E. Glossary of Technical Terms

286, 386           Short for 80286, 80386.

80286, 80386       Microprocessors from Intel used in PCIS.

ASCII              American Standard Code for Information Interchange; a  
  
seven bit representation of alphanumeric characters and  
  
control codes.

ASCII file         A file with ASCII codes; loosely, a text file.

AT                 The name for the second microprocessor generation of  
  
personal computers. These personal computers use the  
  
80286 microprocessor.

Audit trail      A record of changes made to a data set over its lifetime.

Authoring system      Computer software that allows a non-computer programmer to write a CAPI survey questionnaire instrument.

Batch      Computer processing with no human involvement after start-up; the opposite of interactive.

Baud      Baud rate; the number of times per second that a signal in a communications channel changes states; often confused with bps.

Benchmark      The use of some standard computer program (e.g., a sort program) to measure the use of computer resources in a particular environment. This could include computational speed and storage resources.

Bit      Binary digit; symbolically, a one or zero.

bps Bits per second; the number of bits transmitted each second over a communications channel.

Bridge A communications channel between two technically similar networks.

Byte Eight bits.

CAPI Computer Assisted Personal Interviewing is a personal interview usually conducted at the home or business of the respondent using a portable computer.

Case management Portion of the CAPI software that handles the administrative management of the survey. This portion usually includes keeping track of the status of each interview, interviewer assignments, and other similar



administrative tasks.

CASI Computer Assisted Self Interviewing (CASI) involves data collection without the direct presence of an interviewer. CASI can take several different forms which are differentiated by the means of collection. These include Prepared Data Entry (PDE) where the respondent answers questions displayed on a computer terminal; Touchtone Data Entry (TDE) where the respondent answers computer generated questions by pressing buttons on a telephone; and Voice Recognition Entry (VRE) where the respondent answers questions by speaking directly into a telephone.

CASIC Computer Assisted Survey Information Collection.

CATI Computer Assisted Telephone Interviewing (CATI) is a computer assisted survey process which uses the telephone for voice communications between the interviewer and the respondent.

CCITT            Consulting Committee for International Telephony and  
  
Telegraphy; standards setting organization from which  
  
have emerged international standards following their  
  
guidelines in the area of computer networks.

Centralized     Interviews carried out from one central location (e.g.,  
  
nationwide).

Centralized     Main or host computer provides all of the processing  
  
computing power.

Chip            See microchip.

CPU             Central processing unit; the computer part which  
  
interprets and executes instructions.

CRT             Cathode ray tube; the most common type of computer  
  
screen.

Decentralized CATI interviews carried out from several geographically dispersed locations (e.g., states).

97

Distributed processing      Computing power is distributed over a number of computers which may be co-located or geographically distributed.

Disk                      A circular, magnetized medium which holds electronic data.

Disk drive              A device which reads a disk electronically.

Diskette                A floppy disk.

DM                      Direct Manipulation: A type of human-computer interface which accentuates the user's feeling of directly

operating on responsive display objects. Example:

Macintosh user interface.

DOS Disk operating system; an abbreviation for MS-DOS or PC-DOS, the original operating system for IBM PC's.

Download The processing of transferring a file from a mainframe computer or host to a connected personal computer or terminal.

EDI Electronic data interchange; the automated exchange of business information such as invoices.

Establishment Business.

Floppy disk A bendable disk, usually 5 1/4 inches in diameter, although increasing use is being made of unbending disks 3 1/2 inches in diameter.

Gateway A communications channel used to pass data between two

different networks to communicate with each other.

Hard disk      An unbendable disk and its disk drive; holds more data than a floppy disk.

I/O            Input and output.

IDN            Integrated Digital Networks; digital transmission networks which are dedicated to voice and data.

ISDN           Integrated services digital network; an emerging technology which offers many new telecommunication services such as the mixing of the transmission of voice and data.

File server    A computer, usually on a Local Area Network that

provides a group of users with storage facilities to store and access their files.

GB Gigabyte(s).

Gigabyte Loosely, one billion bytes; strictly,  $1,073,741,824$  ( $2$  to the 30th power) bytes.

Interactive Computer processing which prompts for and accepts human input.

KB Kilobyte(s).

Kilobyte Loosely, one thousand bytes; strictly,  $1024$  ( $2$  to the 10th power) bytes.

LAN Local area network; the interconnection of microcomputers at one site.

Mainframe A large computer; often designed to serve many users at

one time, although some mainframes, often called supercomputers, are designed to provide high-speed computing; their purchased costs are often in excess of a million dollars.

MB Megabyte(s).

Megabyte Loosely, one million bytes; strictly, 1,048,576 (2 to the 20th power) bytes.

Microchip A printed circuit etched on a silicon chip.

Microcomputer A small computer, e.g., costing less than \$10,000.

Microprocessor A CPU on a microchip.

Minicomputer A medium sized computer; larger than a microcomputer but smaller than a mainframe; costing on the order of \$100,000.

MS-DOS Microsoft's DOS for PC's.

On-line (1) A peripheral device is on-line when it is connected and ready for use; (2) involving interactive use of a computer.

One-time Non-repeating survey. Data is collected once, or over great intervals (e.g., 5-10 years).

Ongoing Repetitive survey (e.g., weekly, monthly or yearly).

PC Personal computer; broadly speaking, any microcomputer; narrowly speaking, an IBM-compatible computer; even more narrowly speaking, IBM's first microcomputer.

PC-DOS IBM's version of MS-DOS (they are virtually identical).



PDE See Prepared Data Entry

Prepared Data Entry Prepared Data Entry ( PDE) where the respondent answers questions displayed on a computer terminal.

Print server A computer, usually on a Local Area Network that provides a group of users with a range of printing services.

Question path See skip pattern.

RAM Random access memory; the core memory for a computer's CPU.

RAM disk RAM used as if it were disk space.

Sampling Unit A selected element for data collection in a survey' usually selected from a defined population of units by a random mechanism. In a survey of households in a state, the sampling unit is the household.

Skip pattern    The sequence questions are asked in a survey questionnaire instrument; this sequence is often based on the answer to each question.

Target population    Collection of survey units about which you wish to make some measurement, but to quantify it, a sample is obtained and an estimate is calculated.

TDE    See Touchtone Data Entry

Touchtone Data Entry    Touchtone Data Entry (TDE) allows respondents to call and answer questions posed by a computer using the keypad of their touchtone telephone for well-controlled and inexpensive collection.

User-friendly software    Software that provides an interface to the user that is simple and intuitive; thus making the software easily to use.

UNIVAC I        The name of the first digital computer in widespread  
  
commercial use.

100

UNIX            An operatin system initially designed for small  
  
computers, but currently in use over a wide range of  
  
computers.

Upload         The process of transferring a file from a personal  
  
computer or terminal to a mainframe computer or host.

Voice          Voice Recognition Entry (VRE) allows respondents to  
  
Recognition    call and answer questions posed by a computer by  
  
Entry          speaking directly into the telephone. The machine  
  
translates the incoming sounds for verification with  
  
the respondent and storage in a data base.

WAN            Wide Area Network.

Waterfall      A straight-forward approach to software development

methodology by stepping through specification, design, implementation, debugging and testing without ever looking back -- as opposed to moving back and forth between these steps as the objectives become more clearly understood.

WYSIWYG Pronounced whizzy-wig. What You See Is What You Get. A style of presentation to users in which the displayed material is essentially identical in form to the final product. Example: modern word processing software.

XT The name given to IBM to an early version of the Personal Computer which had internal disk storage (i.e., a hard disk) that could hold 10 or more megabytes of data.

Reports Available in the

Statistical Policy

Working Paper Series

1. Report on Statistics for Allocation of Funds (Available through NTIS Document Sales, PB86-211521/AS)
2. Report on- Statistical Disclosure and Disclosure-Avoidance Techniques (NTIS Document Sales, PB86-211539/AS)
3. An Error Profile: Employment as Measured by the Current Population Survey (NTIS Document Sales PB86-214269/AS)
4. Glossary of Nonsampling Error Terms: An Illustration of a Semantic Problem in Statistics (NTIS Document Sales, PB86-211547/AS)
5. Report on Exact and Statistical Matching Techniques (NTIS Document Sales, PB86-215829/AS)
6. Report on Statistical Uses of Administrative Records (NTIS Document Sales, PB86-214285/AS)
7. An Interagency Review of Time-Series Revision Policies (NTIS

Document Sales, PB86-232451/AS)

8. Statistical Interagency Agreements (NTIS Document Sales, PB86-230570/AS)
9. Contracting for Surveys (NTIS Document Sales, PB83-233148)
10. Approaches to Developing Questionnaires (NTIS Document Sales, PB84-105055/AS)
11. A Review of Industry Coding Systems (NTIS Document Sales, PB84-135276)
12. The Role of Telephone Data Collection in Federal Statistics (NTIS Document Sales, PB85-105971)
13. Federal Longitudinal Surveys (NTIS Document Sales, PB86-139730)
14. Workshop on Statistical Uses of Microcomputers in Federal Agencies (NTIS Document Sales, PB87-166393)
15. Quality in Establishment Surveys (NTIS Document Sales, PB88-232921)
16. A Comparative Study of Reporting Units in Selected Employer Data Systems (NTIS Document Sales, PB90-205238)
17. Survey Coverage (NTIS Document Sales, PB90-205246)
18. Data Editing in Federal Statistical Agencies (NTIS Document

Sales, PB90-205253)

19. Computer Assisted Survey Information Collection (NTIS

Document Sales, PB90-205261)

Copies of these working papers may be ordered from NTIS Document

Sales, 5285 Port Royal Road, Springfield, VA 22161 (703) 487-4650