NEW CASIC TECHNOLOGIES AT THE U.S. BUREAU OF THE CENSUS

William L. Nicholls II and Martin V. Appel,
U.S. Bureau of the Census[1]
Washington D.C. 20233-0400

KEY WORDS: CASIC, computer-assisted, data collection, pen computing, disks by mail, touchtone, voice recognition, character recognition, EDI, FAX reporting

1. Background

The U.S. Bureau of the Census is a large, general-purpose, statistical agency. It conducts the Census of Population and Housing in years ending in 0, the Economic and Agricultural Censuses in years ending in 2 and 7, and hundreds of establishment and household surveys on a biannual, annual, monthly, or weekly schedule.

Until quite recently, almost all the Bureau's data collection utilized mail-out, mail-back paper questionnaires or paper interview schedules. This has been changing. In 1992, the Census Bureau established a Computer-Assisted Survey Information Collection (CASIC) Office to "broadly implement CASIC methods in Census Bureau data collection, capture, and processing." (Creighton, Matchett, and Landman 1994).

One function of the CASIC Office is to coordinate the timely and cost effective implementation of two **relatively well proven CASIC technologies**: computer-assisted telephone interviewing (CATI) and computer-assisted personal interviewing (CAPI). To meet these objectives, in 1993 the Census Bureau procured laptop microcomputers to equip most of its field interviewing staff for CAPI and opened additional CATI telephone centers in Tucson, Arizona, and Jeffersonville, Indiana. With the cooperation of the Bureau of Labor Statistics, in January 1994, the Census Bureau moved its largest household survey, the Current Population Survey, to paperless data collection using CATI and CAPI. Under current plans, most Census Bureau household surveys will be fully converted to CAPI or CATI by 1997 and the remainder by the year 2000.

A second function of the CASIC Office is to assess, test, and encourage the implementations of **new and emerging technologies** for possible use in data collection, capture, and processing. The list of technologies to be evaluated is constantly being revised but currently includes:

* Pen-based computing in field operations
* Touchtone data entry
* Voice recognition entry
* Computerized self-administered questionnaires
* Electronic data interchange
* Electronic document imaging
* Optical recognition of hand written characters
* Image processing of FAX data reporting

Not all these technologies are full replacements for traditional paper-based methods. Some represent new methods of capturing data from paper forms. Others may be alternative response modes that the Census Bureau offers to accommodate varying respondent capabilities and preferences.

Many of these same technologies also are being tested and evaluated by other public and private data collection organizations (Blom 1994, Clayton and Werking 1994, Keller 1992 and 1994, Statistical Policy Office 1990, and Tozer and Jaensch, 1994). The Census Bureau's technology evaluation program is unusual, however, in several respects. First, it is systematically examining virtually all major new technologies applicable to survey and census data collection and capture. Second, the testing and evaluation process is viewed as a contribution to general institutional development rather than as a means of meeting the needs of one or more specific surveys. And third, the evaluation process encompasses not only relatively well developed technologies with previously demonstrated cost competitiveness, but also technologies which may not reach this state of maturity until later in the decade.

2. The C^2T^2 Evaluation Process

Direction of the Census Bureau's evaluation of new CASIC technologies was initially assigned to its CASIC Committee for Technology Testing (C^2T^2). Although these responsibilities were transferred this year to the Bureau's CASIC Policy Advisory Group, the three-step evaluation process developed by C^2T^2 is still followed.

The first step is an **initial technical assessment** (ITA) to summarize what is currently known about a candidate technology from publications, formal and informal organizational reports, material provided by vendors, and other easily accessible sources. Each ITA answers a series of standardized questions covering such topics as: range of potential survey and census uses; stage of development; difficulty of application setup; costs of initial investment; user training required; user acceptance; and effects on survey costs, coverage, response rates, estimates, and

timeliness. The ITAs have been written by Census Bureau staff with a primary focus on applications to continuing surveys.

The Census Bureau's Year 2000 Research and Development Staff (2KS), recently merged into the Decennial Management Division, has conducted a closely related program to assess possible data collection technologies for the Year 2000 Decennial Census. While reviewing some of the same technologies and incorporating most of the C^2T^2 assessment criteria, the 2KS program has retained outside specialists in technology assessment to extend the range of examined technologies and to provide details on market trends, technology projections, and likely population penetration by the year 2000.

The second step in the evaluation process is **small-scale feasibility testing** to answer questions unresolved from the ITA, to evaluate the technology's suitability for particular survey applications, or to identify the most appropriate types of hardware and software for production usage. Feasibility testing of promising technologies can occur in a variety of environments, including offices, laboratories, and field tests. The field tests typically employ small, purposeful, or "nonlive" samples, or samples too small to affect the estimates of ongoing statistical series. In some cases, the Census Bureau has underwritten laboratory studies by universities or developed joint evaluation procedures with other Federal agencies. Occasionally, feasibility testing began before completion of an ITA or was already in progress when the C^2T^2 program began.

The third step in the C^2T^2 evaluation is a **large-scale operational test** of the technology in production use. An approved research plan is required for each candidate technology proposed for production use. At a minimum, the test should measure the technology's impact on: survey costs; response rates; data quality; timeliness; and survey estimates. These objectives generally require an experimental design prepared with the assistance of a statistical methods division.

To date, C^2T^2 ITAs (or comparable 2KS reports) have been completed or are in final stages of preparation for a dozen technologies. Ten C^2T^2 or 2KS feasibility tests are complete or nearing completion, while an additional twelve tests are in planning or in progress. Several of the feasibility tests are designed to evolve gradually into production data collection.

3.0 New CASIC Technologies

This section of the paper presents highlights from work to date, drawn both from the technical assessments and feasibility tests and mentions production implementation where applicable.

Pen-based computers use a stylus" or pen for input rather than a keyboard (McGuire and Sebold 1993). The pen creates the illusion of "markings" on the computer screen in one of two ways: (1) by tapping the screen to indicate selections; or (2) by writing information on the screen. The computer recognizes and stores the characters through a "handwriting recognition engine" or records the entry as a "digital ink" image.

Pen computers have several major advantages over keyboard laptop or notebook computers for doorstep **CAPI interviewing** or any entry task performed while standing or outdoors. Because they are designed for mobile field workers, rather than office workers, they: (a) permit one hand entry; (b) are typically of rugged construction; and (c) are intended to operate for long periods on battery power. These advantages have been confirmed by small feasibility tests conducted both by the Census Bureau and by Statistics Canada. The Census Bureau also has tested pen computers for sampling, abstracting, and transmission of discharge data from hospital records and is closely following the development of pen-based CAPI by the Bureau of Labor Statistics (BLS) for the Consumer Price Index Surveys.

Nevertheless, the Census Bureau is unlikely to replace laptop keyboard computers with pen-based computers for general CAPI applications in the next few years. The accuracy of handwritten character recognition is not sufficient for production interviewing and the unit cost of pen computers still exceeds that of keyboard laptops. General-purpose, pen-based CAPI software also is not yet available, although the Berkeley Computer-Assisted Survey Methods Program is developing a pen-based version of the CASES CAI System for BLS. Since the Census Bureau has chosen CASES for its keyboard laptop CAPI applications, this development will facilitate a move to pen-based CAPI when warranted.

The use of pen computers for geographic applications and residential listing is more attractive at present for two reasons (Pfeiffer 1993). First, the graphic capabilities of pen computers are readily adapted to the display, annotation, and updating of maps. Second, pen-based **geographic information systems (GIS)** software is available and in use by public utility companies and others. While the costs of pen-based hardware and software for GIS applications remain high, they are expected to decline before possible use in the Year 2000 Decennial Census.

To gain operating experience with pen-GIS and to assess the ability of current field staff to use it effectively, the Census Bureau field tested a prototype pen-GIS system with 13 pen-based units and 10 field workers in the fall of 1993. This first test was sufficiently promising that four additional tests of pen-based GIS have been scheduled through 1995 to examine enhanced software, to incorporate global positioning system (GPS) capabilities, and to prototype applications in special (local) censuses and the decennial census.

Touchtone data entry (TDE) is an automated data capture technology which allows a respondent, using the keypad of a touchtone telephone, to reply to computer generated prompts (Appel 1992A). The TDE system functions as an interviewer. At a minimum, the system must answer a call, prompt the respondent, recognize touchtone signals, and store the reply.

Survey applications of this technology are limited by technical constraints. First, since a single touchtone key is not unique to an alphabetic letter, responses generally are limited to numbers and multiple choice questions with numeric codes. Second, although an estimated 95 percent of U.S. households have telephones, at least 20 percent of those phones have rotary dials. Even among households with touchtone phones, 20 percent typically refuse to use them for nondialing applications. Third, experience suggests that TDE respondents generally will not answer more than 5 to 10 low-complexity questions per interview. The use of TDE is therefore severely limited for household surveys. The most promising survey applications of TDE have been for brief establishment surveys with regular monthly reporting of numeric data by businesses, industries, or government offices.

At the Census Bureau, TDE is moving from successful prototype testing to the status of a production data collection technology for brief establishment surveys. A central TDE facility is nearing completion at the Census Bureau's Data Preparation Division in Jeffersonville, Indiana. Within a year of its completion, the facility will receive touchtone data from half a dozen Bureau surveys. First production uses will be for the Advanced Monthly Retail Trade Survey and the Quarterly Property Tax Survey.

Voice recognition entry (VRE) is an automated data capture technology which allows a respondent, speaking over a telephone, to reply to computer generated prompts (Appel 1992B). The VRE system functions as an interviewer, reading the questions in a digitized voice, recognizing the respondent's vocal replies, and echoing them back for confirmation. For Census Bureau data collection purposes, only real time, over-the-phone, speaker independent technologies have been considered.

Our initial technical assessment divided VRE technology into three levels by vocabulary size: small vocabulary, limited to the digits 0-9, yes, and no; medium vocabulary of up to 100 words per prompt; and large vocabulary from 100 to thousands of words.

Small vocabulary VRE is a well tested technology best considered an adjunct to touchtone data entry. While its data entry capabilities are similar to those of TDE, small vocabulary VRE can be used with rotary telephones and by respondents who dislike touchtone applications. VRE is more expensive, however. In 1992, TDE PC computer boards cost approximately $400 per phone line while VRE computer boards cost approximately $4000 per line. Like TDE, small vocabulary VRE is primarily useful for short duration, numeric survey applications. Non-numeric responses can be digitized and played back for data keying.

Medium vocabulary VRE is a new technology with promise for (but untested in) survey applications. If a computer-respondent dialogue can be designed such that the respondent's vocabulary is constrained to discrete words, short phrases, and numerics after each prompt, it may be appropriate for medium vocabulary VRE. The Census Bureau has procured a small commercial medium vocabulary VRE system and plans to develop a prototype test survey.

Large vocabulary VRE is an emerging technology that eventually may replace both TDE and simpler forms of VRE. Currently, this technology exists primarily in research laboratories. To explore the potential of large vocabulary VRE, the Census Bureau, through an interagency agreement with the Office of Naval Research, commissioned the Oregon Graduate Institute Center for Spoken Language Understanding and Carnegie Mellon University to build prototype systems for a few decennial census short-form questions. One system used neural-network technology and the second employed hidden-Markov modeling.

First results for the neural network, English-language system with voluntary call-in respondents were very encouraging. Of the more than 4,900 callers who responded to the first prompt, 97.8 percent completed the questionnaire; and recognition of spoken answers exceeded 95 percent to most of the test questions. An expanded prototype, VR call-in system is planned for selected applications in the 1995 Census Test. Major factors in user acceptance are being studied under behavior laboratory conditions by

the Census Bureau's Center for Survey Methods Research.

For **computerized self-administered questionnaires (CSAQ)**, survey agencies send an executable computerized questionnaire (usually on disk) to the respondent who then installs and runs it on his/her own personal computer (Sedivi and Rowe, 1993). No interviewer is present. The automated questionnaire controls the flow of survey questions, provides on-screen instructions, and may include edit checks. Some systems allow the user to import historic data. The respondent returns the answered disk by mail or transmits the data by modem. This survey method is sometimes known as the "disks by mail" technique (Pilon and Craig 1988) or "prepared data entry (PDE)" (Statistical Policy Office 1990).

Use of CSAQ is limited by the penetration of PCs within target populations. At present, only an estimated 26 percent of U.S. households have PCs, and this percentage is not expected to increase beyond 40 percent through the end of the decade (Ogden Government Services 1993). CSAQ is not, therefore, a promising survey method for the general public, except perhaps when offered as one of several options the public could choose from to respond to a large survey or census. Personal computers are much more prevalent in business. An estimated 98 percent businesses, and an estimated 63 percent of small businesses (with 100 or fewer employees), have PCs. Successful governmental users of this method have either: confined their study populations to respondents known to have compatible PCs; or prescreened potential respondents and limited CSAQ participation to those with the appropriate PC environment. Several Federal agencies have reported successful uses of CSAQ: the Petroleum Supply Division of the Energy Information Administration (Statistical Policy Office 1990); the National Center for Education Statistics (Kindell 1992); and the National Science Foundation.

The Census Bureau has successfully completed one internal feasibility test of CSAQ, begun small-sample test production for the Company Organization Survey, and is planning further production tests of CSAQ in three additional establishment surveys: the 1994 Annual Survey of Manufactures; the 1994 Survey of Industrial Research and Development; and the 1995 Monthly Fats and Oils Survey.

The Census Bureau's current CSAQ development efforts are focussed on: use of computer bulletin boards systems for respondent access to CSAQ instruments and return of data; incorporation of off-the-shelf forms design software with graphical user interfaces for design of CSAQ instruments; and enabling respondents to link data in their computer files or databases with response fields in CSAQ questionnaires.

Electronic data interchange (EDI) is the electronic transfer of business transaction information in a standard format between business partners (Ambler and Messenbourg 1992). In U.S. business applications, EDI transactions usually are computer-to-computer over a value added network using the American Standard (X12) transaction set. An international standard, the UN/EDIFACT message format, also is gaining worldwide acceptance. Large companies are increasingly replacing exchanges of paper forms with EDI for such everyday business transactions as buying, shipping, selling, billing, and inventory control. The largest companies have encouraged the Census Bureau to consider EDI for reporting data to the Economic Censuses and to continuing surveys.

The Census Bureau has received data electronically from selected respondents for many years. The Census Bureau's Foreign Trade Division abstracts and summarizes U.S. import transactions that import brokers report to the U.S. Customs Service's Automated Broker Interface. The Census Bureau's Governments Division receives much of its financial data on school districts, local, and State governments on magnetic data tapes and diskettes. Very large companies also have responded to the Economic Census with data tapes. These data sets have relied on function- or survey-specific transaction sets.

The new Census Bureau emphasis is to encourage data reporting using officially approved EDI standards. Transaction Set 152, has been developed and approved by the ACS X12 Committee of the American National Standards Institute (ANSI) for the reporting of economic statistical information to the Census Bureau. The Census Bureau also is working with EUROSTAT to propose a similar statistical reporting transaction set to the UN/EDIFACT Board in 1995. Reporting procedures using the X12 standard were first tested with one large company of more than 500 establishments in the 1992 Economic Census. Agreements also have been reached with 10 large companies for X12 EDI reporting on the Company Organization Survey for the 1993 reporting year. A continued increase in direct X12 EDI reporting is anticipated in the future for these and other economic surveys.

The Census Bureau also has developed a CSAQ data collection package which includes backend software modules to convert entered data into an appropriate EDI format (e.g., X12 Set 152) and to facilitate the transmission of that formatted data to

the Census Bureau. The CSAQ instrument can be completed by staff without programming skills, so that responding companies may submit data in appropriate EDI format without devoting their limited EDI systems staff to the conversion process.

Electronic document imaging refers to a process in which paper documents, such as survey or census forms, are electronically scanned, converted to digital images and then stored in computer retrievable form. Subsequent processing operates on the form's electronic image, or data abstracted from it, rather than from the original paper document. Electronic imaging is now used extensively in place of microfilming to store, retrieve, display, and print reference works, periodicals, business forms, and governmental records. The Internal Revenue Service and the Patent and Trademark Office are migrating to more extensive use of electronic imaging for such functions.

An initial technical assessment, completed by Synectrics for Management Decisions in January 1994, recommended that the Census Bureau move as quickly as possible to acquire and pilot test a data capture system using the latest imaging and recognition technology. The Census Bureau subsequently entered into a multi-year research and development project with the Rochester Institute of Technology Research Corporation to develop an advanced document scanning system for the Year 2000 Decennial Census. A second vendor team is being assembled to integrate recognition software into the system. Pilot tests are planned for the 1995 Census Test.

Optical character recognition (OCR) is a methodology in which computer software recognizes each element of an alphanumeric character string and converts it to the corresponding ASCII code, e.g., the image of the letter "A" is converted to the ASCII code for "A". OCR systems are now commonly used to convert libraries of machine-printed documents (e.g., books, magazines, records) into ASCII format to allow indexing and text searching. Advanced OCR systems are now being tested to capture hand written answers to census and survey forms, thereby reducing the direct data entry workload. The Census Bureau is exploring OCR as part of its evaluation of potential technologies for the Year 2000 Decennial Census. Survey applications have been reported at Statistics Sweden (Blom 1994) and at the Australian Bureau of Statistics (Tozer and Jaensch 1994)

In May 1992 and in February 1994, the Census Bureau hosted two Census OCR Systems Conferences with the National Institute of Standards and Technology (NIST) (Wilkinson *et al.* 1992 and Geist *et al.* 1994). OCR vendors representing significant portions of the industry were given large, standard image test sets on CD-ROM to process. The results were scored, tabulated, and reported by NIST at conferences where the vendors were invited to describe their methods.

The first test set consisted of separated numeric and alphabetic characters hand-printed by Census Bureau employees and by high school students from Montgomery County, Maryland. About half the participating firms recognized about 95% of the digits, 90% of the upper case letters, and over 80% of the lower case letters. Accuracy near 100% is not essential for production data capture if the software flags characters whose identification confidence falls below a preset threshold. Clerical staff may then employ human recognition to enter values for the flagged images.

The second test set consisted of images of hand written entries to selected questions (including reports of occupation and industry) from a national probability sample of the 1990 Census long forms. The paper forms were double keyed (and triple keyed for mismatches) to establish a "truth" reference set. The performance of the automated OCR systems was compared with that of human production census keying. One OCR system achieved accuracy comparable to human production keying for about half the fields while rejecting the remaining half for operator correction. Advances in recognition were primarily a function of two factors: better segmentation of characters or particles; and matching of full fields against dictionaries of words and phrases. Optical character recognition will be production tested for similar tasks in processing of the 1995 Census Test.

Image processing of FAX data returns (IP-FDR) represents an extension of imaging and OCR technologies (Appel and Rowe 1993). At the Census Bureau, increasing numbers of establishment survey questionnaires and report forms are being returned by facsimile (FAX) equipment rather than by mail. This change has been initiated in part by respondents who viewed FAX data reporting as simpler, faster, or more convenient than mailing returns. Survey managers have encouraged this trend by establishing toll free 800 FAX numbers.

Currently, paper copies generated by the FAX machines are combined with mail returns for clerical key entry. An alternative is to process the FAX electronic image using OCR methods, although to our knowledge this approach has not previously been tested for survey applications. To explore this

alternative, the Census Bureau has constructed a small prototype IP-FDR system from commercially available products. This system can: receive FAX transmitted forms through the public telephone network 24 hours a day; identify the survey and respondent to which the form pertains; utilize mark reading capabilities for checked boxes and filled circles; employ OCR capabilities to recognize machine-printed and hand-printed alphanumeric characters in prespecified fields; and display all or a portion of the imaged document on a computer terminal for adjudication or key entry when necessary. In an initial feasibility test, 50 respondents to the Manufacturers' Shipments, Orders, and Inventories Survey (M3) were asked to return their monthly reports to the prototype system by FAX. During February and March of 1994, with the IP-FDR system's OCR confidence level set at 100 percent, 60 percent of the characters were recognized and 70 percent of the fields were completely correct. A more sophisticated OCR recognizer should increase these percentages and reduce the proportion of forms requiring clerical adjudication and entry in at least one of the fields. System enhancements and procedural refinements are continuing for the M3 test while operational testing with additional economic surveys is planned.

4. Discussion and Conclusions

This paper has presented an overview of eight new technologies which show promise for census and survey applications in the near and more distant future. Several caveats seem necessary to stress the difficulty of drawing enduring conclusions in this rapidly evolving area.

First, it should be recognized that the information presented here is subject to revision as technological capabilities increase and additional experience is gained in their survey and census applications. This paper can only purport to summarize information known about them at a given point in time. At least some statements included here will surely be outdated in the near future. Second, even the list of technologies worthy of careful assessment is subject to continual change. This paper reviews only eight technologies on which there is relatively stable staff consensus of potential importance. The list of candidate technologies proposed for current or future ITAs or testing (or at least close monitoring over time) is twice as long. Third, several of the reviewed technologies clearly have the potential of merging into one another. Touchtone and small vocabulary voice recognition data entry often are used together, and the same telephone receiving facility may be used to accommodate data collected by TDE, VRE, CSAQ, and IP-FDR. Document imaging, optical character recognition, and image processing of FAX data returns must be closely integrated for maximum efficiency. As previously mentioned, the Census Bureau is developing the CSAQ functions to convert responses to X12 formats for EDI transmissions. The combination of pen-based CAPI with voice technologies represents another promising area of future development.

At the same time, a few general cautions about the general applicability and current cost efficiency of these technologies seem necessary.

First, applications of many of these technologies are currently limited by constraints on the type and amount of data they can readily collect from respondents. Other applications are limited by respondent access to specialized equipment or the knowledge they require, such as touchtone telephones, FAX equipment, and personal computers and EDI standards. None of these technologies are general-purpose survey data collections methods at present. Second, many of these technologies are currently most cost efficient in long-term, large-scale use. At present, the costs of initial hardware acquisition and survey setup may preclude the use of pen CAPI and GIS, large vocabulary VRE, EDI, imaging, OCR, and IP-FDR for small or frequently changing survey applications.

Finally, since the growth of some of these technologies is so rapid, prudent managers may wish to defer heavy investments in them until appropriate standards are in place to ensure extended use of hardware and software purchases.

REFERENCES

A list of meaningful references is too extensive to include. If you would like a copy, please contact: William Nicholls, Bureau of the Census, CASIC, Washington DC 20233.