NEW CASIC TECHNOLOGIES AT THE U.S. CENSUS BUREAU

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

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

1. Background

The U.S. Census Bureau 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 is 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."

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, the Census Bureau has recently procured 2,160 laptop microcomputers to equip the majority of its field interviewing staff for CAPI and opened a second CATI telephone center in Tucson, Arizona. With the cooperation of the Bureau of Labor Statistics, the Census Bureau is moving its largest household survey, the Current Population Survey, to paperless data collection using CATI and CAPI in January 1994. Under current plans, most Census Bureau household surveys will be converted to CAPI or CATI by 1996 and the remainder by the year 2000.

A second function of the CASIC Office is to assess, test, and evaluate **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 includes:

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

These new technologies are not necessarily replacements for traditional paper-based methods. Some represent new methods of capturing data entered on paper forms. Others may be used as alternatives to traditional data collection methods for more technically sophisticated respondents to establishment surveys or censuses. Provision of alternative means of replying to survey and census requests may be one means of accommodating varying respondent preferences.

Many of these same technologies also are being tested and evaluated by other public and private data collection organizations (Blom 1993, Keller 1992, Statistical Policy Office 1990, Werking, Tupek, and Clayton 1988). 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 but in time for use in the year 2000 Census.

2. The C^2T^2 Evaluation Process

Direction of the Census Bureau's evaluation of new CASIC technologies has been assigned to its CASIC Committee for Technology Testing, (C^2T^2), which has established a three-step evaluation process. Each step is to be carefully documented.

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.

Promising technologies proceed to **small-scale feasibility testing** to answer questions unresolved from the ITA, to evaluate their suitability for particular survey

applications, or to identify the most appropriate types of hardware and software for production usage. Feasibility testing 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 another Federal agency. Occasionally, feasibility testing begins 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.

To date, seven ITAs have been completed or are in final stages of preparation. About an equal number of feasibility tests are in progress or in planning. Two of the feasibility tests will evolve into operational tests. None has yet been completed.

3.0 New CASIC Technologies

This section of the paper presents highlights from work to date, drawn both from the initial technical assessments and from the feasibility tests.

Pen-based computers are defined as computers that provide users with 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 in "digital ink." Because pen-based computers are designed for mobile field workers, rather than office workers, they: (a) permit one hand entry; (b) typically are of rugged construction; and (c) are designed to operate for long periods on battery power.

Pen computers have several major advantages over laptop or notebook computers with keyboards for doorstep **CAPI interviewing** or any entry task performed while standing or outdoors. These advantages have been confirmed by small feasibility tests conducted both by the Census Bureau and by Statistics Canada. Nevertheless, pen-based computers are unlikely to replace laptop keyboard computers for CAPI in the near future. None of the competing pen-centric hardware designs or operating systems have yet become a dependable standard; handwritten character recognition engines are not sufficiently reliable; general-purpose, pen-based CAPI software is not yet available; and the unit cost of pen computers exceeds that of laptops with keyboards. Although some experts believe these problems will be solved within the next few years, heavy investments in this highly volatile field seem risky.

The use of pen computers for residential listing and **field operations of geographic information systems (GIS)** 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 GIS software already exists and is in use by public utility companies and others. While the costs of pen-based GIS hardware and software are high, they are expected to decline before this technology would be broadly used in the next census. To gain operating experience with pen-GIS and to assess the ability of Census Bureau field staff to use it effectively, the Census Bureau is beginning small tests of prototype pen-GIS systems. If initial tests seem promising, larger tests may be undertaken in the 1995 Census Test.

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, reading the questions and echoing the touchtone entries in a digitized voice. 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 each alphabetic letter, responses generally are limited to numbers and multiple choice questions with numeric codes. Second, although an estimated 94 percent of U.S. households have telephones, at least 35 percent have rotary dials. Even among households with touchtone phones, 20 percent typically refuse to use them for nondialing applications. 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. This use was pioneered by the U.S. Bureau of Labor Statistics (Werking, Tupek, and Clayton 1988). The Census Bureau is now testing (or preparing to test) TDE in three establishment surveys: the Manufacturers'

Shipments, Orders, and Inventories Survey, the Monthly Advanced Retail Trade Survey, and the Survey of Construction. Most businesses and industries have touchtone phones, and many find TDE more convenient for monthly statistical reporting to Federal agencies than mailed forms. Offering monthly survey reporters a TDE option does not encourage reluctant reporters to respond; but TDE does increase reporting convenience and timeliness for cooperative respondents and reduces survey costs. For respondents who do not have touchtone phones, a voice recognition backup can be used.

At the Census Bureau, TDE is currently moving to the status of a production technology. A central TDE receiving facility is being prepared at the Census Bureau's Data Preparation Division in Jeffersonville, Indiana, to receive touchtone data from all Bureau surveys employing this technology. In time, this facility will be enhanced to accommodate the receipt of additional telephone-based data collection technologies.

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.

It is convenient to divide VRE technology into three levels by vocabulary size: (1) small vocabulary, limited to the digits 0-9, yes, and no; (2) medium vocabulary of up to 100 words per prompt; and (3) large vocabulary from 100 to thousands of words. For Census Bureau data collection purposes, only real time, over-the-phone, speaker-independent technologies were considered.

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 commercial medium vocabulary VRE system and is developing 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, has 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. The first system will use neural-network technology and the second will use hidden-Markov modeling.

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, Statistical Policy Office 1990). No interviewer is present. The automated questionnaire controls the flow of survey questions, provides on-screen instructions, and may include edit checks (and items to reconcile edit failures) performed as the data are entered. Some systems allow the user to import data from Lotus, dBASE, or similar files. 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 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 of large businesses, and an estimated 63 percent of small businesses (with 100 or fewer employees), have PCs. Successful governmental users of this method have either: (1) confined their study populations to respondents known to have compatible PCs and the knowledge to operate them; or (2) prescreened potential respondents and limited CSAQ participation to those with the appropriate PC environment and knowledge. Both the Petroleum Supply Division of the Energy Information Administration (Statistical Policy Office 1990) and the

National Center for Education Statistics (Kindell 1992) have reported successful uses of CSAQ. The Census Bureau is planning feasibility tests of CSAQ for two establishment surveys in 1994, the Company Organization Survey and the Annual Survey of Manufactures. This new method also is under consideration for three additional Census Bureau establishment surveys.

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, that transfer usually is computer-to-computer over a value added network using the American Standard (X12) transaction set. An international standard, the UN/EDIFACT message format, also exists. 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 data reporting to the Economic Censuses and to continuing surveys.

The Census Bureau has received data electronically from selected respondents for many years. Ninety-one percent of U.S. import transactions are reported to the U.S. Customs Service by import brokers using an Automated Broker Interface. After the data are processed through Census Bureau and Custom Service edits, they are abstracted and summarized by the Census Bureau's Foreign Trade Division. The Census Bureau's Governments Division also 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-specific transaction sets (as with foreign trade) or on survey-specific, detailed specification for data tapes. The latter often have proven time consuming and cumbersome to employ both for respondents and for the Census Bureau.

The new Census Bureau emphasis is to encourage the reporting of data using officially approved EDI standards. A new official X12 standard for the reporting of economic statistical information to the Census Bureau (Transaction Set 152) was developed and approved by the ACS X12 Committee of the American National Standards Institute (ANSI). This transaction set incorporates many standard X12 data elements and segments to facilitate the preparation of reports by companies already utilizing X12 EDI in their business transactions. These new EDI reporting procedures were first tested with two large companies in the 1992 Economic Census. The Census Bureau is now proceeding to implement EDI in its Company Organization Survey and the Annual Survey of Manufactures (Ambler and Hyman forthcoming)..

Optical imaging, or more precisely, "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 readable form. Subsequent processing operates on the form's electronic image, or data abstracted from it, rather than from the original paper document. Optical imaging is now used extensively to store, retrieve, display, and print reference works, periodicals, business forms, and governmental records. Optical imaging is increasingly used as a replacement for microfilming or microfiching of documents.

In census and survey applications, imaging may be used for capture, transmission, storage, archiving, and retrieving of survey and census forms. Images can be displayed on computer screens for key entry or for review by analysts.

The Census Bureau currently employs microfilming for many of its storage, archiving, and edit-reference functions; but it is considering optical imaging as an alternative. Key considerations may include: (a) relative costs of imaging and microfilming; (b) accessibility of records at various processing stages; (c) the probable stability of current imaging standards into future decades; and (d) the prospects for efficient character recognition of hand-written entries from those images.

Optical character recognition (OCR) is a data capture method in which computer software recognizes each element of an alphanumeric character string and converts it to the corresponding ASCII code. 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. Future OCR systems may automate the capture of hand printed and hand written answers to census and survey forms, thereby reducing the direct data entry workload. Survey applications observed at Statistics Sweden (Blom 1993), and administrative uses reported by the Wyoming Department of Revenue (Appel and Rowe 1993) show promise.

In May 1990, as part of its evaluation of potential technologies for the year 2000 census, the Census Bureau and the National Institute of Standards and Technology (NIST) hosted the First Census OCR Conference (Wilkinson *et al.* 1992). OCR vendors representing a significant portion of the industry were given a large, standard image file on CD ROM to process. The test consisted of separated numeric and

alphabetic characters hand-printed by Census Bureau employees. NIST tabulated the results and found that 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 OCR software identifies and flags unrecognized characters when their identification probability falls below a preset threshold. Cases with flagged characters may be displayed at computer terminals (or PCs) where clerical staff use human recognition to make entries for the flagged (highlighted) characters. The accuracy of the combined process has been found satisfactory in limited testing at Statistics Sweden.

A second Census-NIST joint assessment of current OCR capabilities, and a second Census OCR Conference, is planned to extend this work. This evaluation will test the ability of vendors' current optical character recognition (OCR) or intelligent character recognition (ICR) software to recognize a mixture of hand printed and hand written (cursive) entries drawn from selected questions of the 1990 Census. The entries will be taken from a national sample of microfilmed Census records. First results should be available in early 1994.

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. In part, this change was initiated 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 and batch editing. An important alternative is to process the FAX electronic image rather than its paper copy.

Commercial products can be integrated to produce a system for image processing of FAX transmitted forms which: (1) are available through the public telephone network 24 hours a day; (2) are capable of identifying the survey and respondent to which each appropriately coded questionnaire pertains; (3) can display all or a portion of the imaged document on a computer terminal for adjudication or key entry; (4) have OCR capabilities for machine-printed and hand-printed alphanumeric characters; and (5) have mark reading capabilities for checked boxes and filled circles. The Census Bureau is constructing a small prototype system and will test it on three survey instruments of differing style.

4. Discussion and Conclusions

This paper has presented an overview of eight technologies which show promise for survey and census applications in the near and more distant future. Several caveats seem necessary to stress the difficulty of drawing enduring, firm conclusions about 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 working list of candidate technologies proposed for current or future ITAs or testing (or at least close monitoring over time) is twice as long . For example, an initial technical assessment on multimedia computing (Bell *et al.* forthcoming) is nearing completion.

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. One Census Bureau division has proposed the use of CSAQ (disks by mail) to assist respondents in preparing data in X12 formats for EDI transmissions. The combination of pen-based CAPI with voice technologies represents another promising area of development for the future.

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 may be limited by respondent access to specialized equipment or knowledge they require, such as touchtone telephones, FAX equipment, personal computers, and EDI standards. None of these technologies are general purpose survey methods at present.

Second, many of these technologies are currently most cost efficient in long-term, large-scale use. The costs of initial hardware acquisition and survey setup may currently preclude the use of pen CAPI and GIS, larger vocabulary VRE, EDI, imaging, OCR, and IP-FDR for small or frequently changing survey applications.

Finally, the rapid growth of some of these technologies may discourage prudent survey managers from investing heavily in hardware and software which may become obsolete within a year or two. It is not sufficient to know what a technology can do and its current procurement and development costs. Prudent managers also need to know when a technology has sufficiently stabilized to justify major investments for an extended period of years.

---

[1]This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

REFERENCES

Ambler, Carole, and Messenbourg, Thomas L. (1992). EDI - Reporting Standard or the Future, *Proceedings of the Bureau of the Census 1992 Annual Research Conference*, 289-297.

Ambler, Carole and Hyman, Stanley (forthcoming). Electronic Data Interchange: Initial Technical Assessment. A report to the CASIC Committee for Technology Testing, U.S. Census Bureau.

Appel, Martin V. (1992A), Touchtone Data Entry: Initial Technical Assessment. A report to the CASIC Committee on Technology Testing, U.S. Census Bureau.

Appel, Martin V. (1992B). Voice Recognition Entry: Initial Technical Assessment. A report to the CASIC Committee on Technology Testing, U.S. Census Bureau.

Bell, J. Jerry; DeFazio, Karen H.; Curtis, Edward S.; Donaldson, Lyn A. (forthcoming). Multimedia Technology: Initial Technical Assessment. A report to the CASIC Committee on Technology Testing, U.S. Census Bureau.

Blom, Evert (1993). New Technologies in Data Collection at Statistics Sweden. Paper given at the 1993 Field Technologies Conference.

Keller, Wouter J. (1992) Electronic Data Processing in Official Statistics. Paper presented at the First International BLAISE Users Meeting.

Kindell, Carrol Brenner (1992). Electronic Data Collection at the National Center for Education Statistics: Successes in the Collection of Library Data. *Proceedings of the Bureau of the Census 1992 Annual Research Conference*, 298-310.

McGuire, Patty and Sebold, Janice (1993). Pen Com puting Technology: Initial Technical Assessment. A report to the CASIC Committee on Technology Testing, U.S. Census Bureau.

Ogden Government Services (1993). *U.S. Bureau of the Census Technology Assessment of Data Collection Technologies for the Year 2000: Final Technology Assessment Report*. A report prepared for the U.S. Bureau of the Census Year 2000 Staff.

Pfeiffer, Alfred (1993). Pen Computer Field Geographic Information SystemsInitial Technical Assessment. A report to the CASIC Committee on Technology Testing, U.S. Bureau of the Census.

Pilon, Thomas L. and Craig, Norris C. (1988). Disks-by-Mail: A New Survey Modality. Paper presented at the 1988 SawTooth Software Conference.

Rowe, Errol and Appel, Martin V. (1993). Image Processing of Facsimile Data Reporting: Initial Technical Assessment. A report to the CASIC Committee on Technology Testing, U.S. Bureau of the Census.

Sedivi, Barbara and Rowe, Errol (1993). Computerized Self Administered Questionnaires by Mail or Modem: Initial Technical Assessment. A report to the CASIC Committee on Technology Testing, U.S. Bureau of the Census.

Statistical Policy Office, U.S. Office of Management and Budget (1990). *Computer Assisted Survey Information Collection*. Statistical Policy Working Paper 19.

Werking, George S., Tupek, Alan, and Clayton, Richard (1988). CATI and Touchtone Self-Response Applications for Establishment Survey, *Journal of Official Statistics*, Vol. 4, 349-362.

Wilkinson, R. Allen; Geist, Jon; Janet, Stanley; Grother, Patrick J.; Burges, Christopher J.C.; Creecy, Robert; HammondBob; Hull, Jonathan J.; Larsen, Norman J.; Vogl, Thomas P.; and Wilson, Charles L. (1992) The First Census Optical Character Recognition System Conference. A report of the National Institute for Standards and Technology.