LARGE SURVEYS ON SMALL COMPUTERS: NCHS' CAPI SYSTEM

John S. Gardenier, Centers for Disease Control,
National Center for Health Statistics,
6525 Belcrest Road, Room 1120, Hyattsville, MD 20782

Key Words: CASIC, CAPI, Computers

CASIC, CATI and CAPI at NCHS

As computer capabilities have developed, federal statistical agencies have explored how automation can best help them streamline large national surveys. The Office of Management and Budget was interested enough in the potential advantages to convene a subcommittee of the Federal Committee on Statistical Methodology. The subcommittee reported in May, 1990 that computer-assisted survey information collection (CASIC) can improve the quality, timeliness, and flexibility of statistical data gathering.[1]

The National Center for Health Statistics (NCHS) conducted early experiments on Computer Assisted Telephone Interviewing (CATI), especially in considering ways to reduce the cost of the National Health Interview Survey (NHIS).[2] The lower response rates and the lack of universal telephone coverage led the Center to decide that personal household interviews were still the preferred method for its population-based surveys. These surveys are so large and complex that skilled on-site interviewers remain necessary.

That is not to say that NCHS does not rely on CATI. It does, but only for following people who have already been interviewed in person.

The preference for personal interviewing led NCHS to another form of survey automation - the computer assisted personal interview (CAPI). CATI can run on a mainframe or minicomputer, at a central location, with direct supervision over interviewers. It does not have to contend with severe limitations on memory size or processing speed. In contrast, CAPI must be implemented on small, lightweight computers that can be carried into the field. That means CAPI systems have to be logically complete and highly reliable. Supervision tends to be remote both in space and time. That means CAPI systems must be flexible enough to allow the interviewer to deal with any real world interviewing scenario that may arise. NCHS tried several different approaches before arriving at a decision on how to automate personal interview surveys.

In 1987, the NCHS, aided by U. S. Bureau of the Census interviewers, conducted a feasibility study of NHIS automation to determine how interviewers and respondents would react to using computers for household surveys.[3] The technology of that time was not sufficient to implement the full NHIS. Still, the results were so encouraging that the NCHS purchased laptop computers and used them for quick turnaround data gathering about Acquired Immune Deficiency Syndrome (AIDS) knowledge and attitudes in 1988 and 1989. In 1990, the same technology combined the AIDS with another NHIS supplement on Health Promotion and Disease Prevention.

In 1988, the Third National Health and Nutrition Examination Survey (NHANES III) was going into the field. The same issues confronted this survey. The designers were well aware that it was essential to automate as much as possible if the data were to be timely. Many of the examinations and interviews administered in the mobile

examination centers were automated. Because of the large instrument/small computer problem, the household interviews were still conducted by pencil and paper.

In all cases, the interviews were individually programmed, generally in a microcomputer-oriented database language.

Automation Strategies

Often, people adapt existing CATI systems to the CAPI task. Most CATI systems have been designed for mainframes or minicomputers. In such systems, there is generally no problem with memory size, the power supply, or the speed of operation. All of these problems can arise with the small computers used in CAPI systems.

Also, software and data architectures well-suited to larger computers do not necessarily translate well to microcomputers.

For example, CATI systems usually have a fixed size, rectangular file structure. If the survey addresses variable numbers of persons and varying numbers of questions depending on characteristics of each person, then the file structure has to be set for the maximum anticipated response file size. For the bulk of the interviews, the actual demand for file space will be very much smaller than the fixed structure. The suitability of this architecture for laptops depends on the rectangular file size, which is a function of the number and length of the variables.

Trying to implement very large, complex surveys on laptop computers leads to consideration of a second approach. That is, designing the CAPI system from scratch to meet survey requirements and the target computer characteristics.

CAPI systems must be implemented on portable, preferably laptop, computers. The computers have to be small and light enough to be carried easily in field data gathering. (Although field interviews are typically performed inside peoples' homes, on occasion they are done on porches, car or tractor fenders, in gardens, or in farm fields. The environments may run from households with full temperature control and excellent lighting to outdoor settings in extremes of heat or cold and glare or dimness.) The laptop computers have to operate on batteries where AC power is not available. On some days, interviewers must conduct a succession of interviews relying on batteries. If these surveys run from two to four hours each, battery power conservation is especially important.

One way to conserve battery power is to minimize reliance on disk input and output. That can be done by including sufficient memory in the computer to allow use of surrogate disks in random access memory (RAM). A side benefit of that strategy is that interview software runs faster by avoiding disk input and output. The first generation NCHS CAPI laptop had no hard disk. It loaded the interview software from floppy disk, saved the response data in RAM disk, and only wrote back to the floppy at the end of the interview. This usually worked well, but it risked loss of considerable data if the interview were interrupted accidently. That led to a design goal of writing to disk at intervals such that the risk of data loss is acceptable.

Another important consideration in designing a CAPI system from the ground up is the need for speed and smooth flow. The system must aim for assured data quality rather than speed at all costs. Still, awkward interruptions or hesitations present

opportunities to break off. The interviewers' prefer a fast, smooth session. Supporting that preference is likely to improve response rates.

Yet another consideration in small computers is the relatively slow speed of disk access. This problem is best controlled by minimizing the disk accesses and the scope of disk searches, as well as using a virtual disk in RAM memory rather than a physical disk for most processing. The file size should be adaptable to the actual needs of a given survey in a given household. That calls for dynamic adaptation of the file structure to each interview and argues against fixed rectangular formats.

One final note: because of the poor lighting conditions which are not uncommon in field settings, screens have to be designed for minimum clutter and maximum readability.

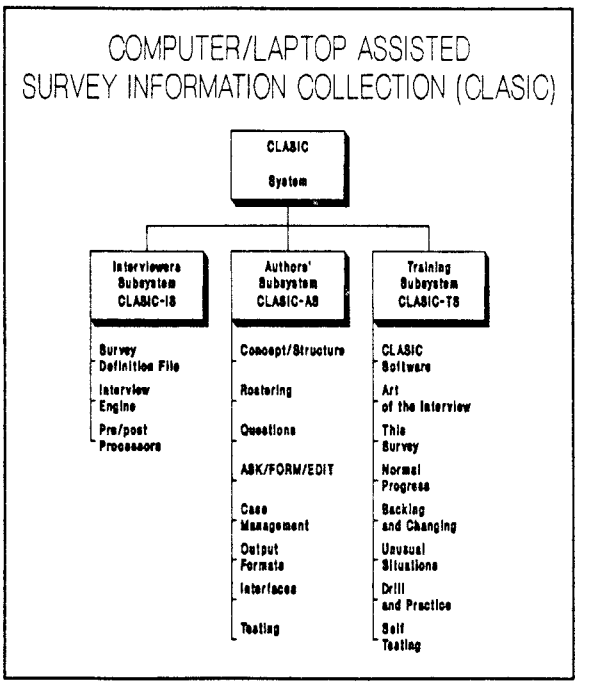
The CLASIC Software System Concept

There are several excellent survey software systems in use in the United States and other countries. These systems have registered impressive successes in market research, demographics, and labor force surveys. Before proceeding into development, we tested many of them. Unfortunately, not one met all of the generic needs summarized above. The reason is largely that the NCHS health interview surveys are more complex and variable than most other surveys. It makes sense to meet these unique needs by designing a system specifically for this class of surveys.

Based on the needs of the various surveys, the NCHS CAPI staff developed and published a concept for a comprehensive software system.[4] The title was later generalized to account for record data entry, institutional interviewing, and telephone interviewing within field surveys. Today, a classier acronym has been adopted, Computer/Laptop Assisted Survey Information Collection (CLASIC). As shown in Figure 1., the initial subsystems to be developed are:

CLASIC-IS Interviewers' Subsystem,
CLASIC-AS Survey Authors' Subsystem,
CLASIC-TS Interviewer Training Subsystem.

Figure 1. CLASIC System Concept



The Interviewers' Subsystem is the key. It will be used directly to conduct interviews. In the Survey Authors' Subsystem, CLASIC-IS becomes the means for testing elements of a survey during its development. It will also be useful within a generic computer-based instructional system for teaching interviewers how to conduct specific surveys.

We were careful to avoid basing the requirements on any single survey. Rather, the CAPI staff documented the needs of all major NCHS interview surveys, as reflected in existing paper instruments. Specialists from each survey program reviewed and commented on those requirements.

The NCHS CLASIC-IS Software Goals for the Interviewers' Subsystem

The software requirements specifications for the first major subsystem, the Interviewers' Subsystem are quite extensive. The key goals are summarized in Table 1.

Table 1. CLASIC-IS Key Goals

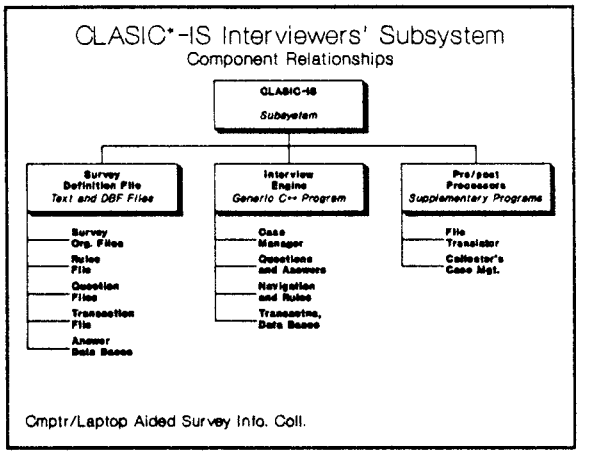
OPERATIONAL GOALS FOR CLASIC-IS
• Generic Software with Strong Maintainability, Reusability
• Single Program to Run All NCHS CAPI Surveys (Of Any Size) on Laptop Computers
• Survey Text and Logic in Supplemental Files
• Ability to Back up and Change Answers Without Data Loss or Corruption
• Ability to Resume Interrupted Interview At or Near Point of Interrupt
• Allowance for Multi-Level Dynamic Rostering
• Keeping Score on What Is Needed/Done in Interview
• Reformatting Output Files to User Specifications
• Interfacing with Different Case Management Systems

The detailed requirements specifications and the design overview were published in an NCHS working paper in January, 1991.[5]

The CLASIC-IS Design and Results

The overall design which successfully meets the goals is summarized in Figure 2. This system is the only one we know of which is currently capable of implementing the full NHIS. That is not to say it has every capability survey designers could wish for. For example, it cannot now support randomizing multiple choice answers. That and other capabilities will be added in future versions.

Figure 2. CLASIC-IS Design Overview



We attempted, unsuccessfully, to prepare CLASIC-IS within sufficient lead time to test, train, and field the 1992 NHIS. The major problem was the shear complexity of the NHIS core instrument. The implementation has shifted to 1993. As of August, 1991, the software and the NHIS core were in initial beta test.

All detailed requirements specifications had included explicit test criteria to aid in developing verification and validation test plans.

We are working to add three elements to the NHIS implementation:

1. The interface with a new automated survey management system being developed at the Bureau of the Census, the data collection agency.
2. The postprocessor module to translate the answers to a common interchange format for use by mainframe computers.
3. The 1993 NHIS supplements.

The new CLASIC-IS system will have to support at least two major surveys. Otherwise, it could not be shown to be generic. Given the NHIS delay, the NHANES interviews may be the first CLASIC-IS surveys fielded.

More surveys will be implemented later. Ultimately, all NCHS interview surveys should migrate to the CLASIC system because of its ease of revision, update, maintenance, and configuration control.

Conclusion

The decision to embark on a major new software development process for statistical survey data collection should never be made lightly. It is difficult, expensive, and almost certain to leave some user expectations unfulfilled. The best way to make that decision is after a thorough analysis of the absolute requirements of (not desires for) the surveys to be supported, and careful evaluation of existing software.

If decision is to proceed with a new development, then the paradigms and principles of modern software engineering should be employed. Requirements should be specified to point to some practical way of testing whether the developed software meets that requirement. Ability to state the test helps, up front, to validate the legitimacy of the requirement. It also aids development of a test plan.

The author is grateful to Mary Grace Kovar, NCHS, for reviewing this article.

1. Subcommittee on Computer Assisted Survey Information Collection of the Federal Committee on Statistical Methodology, "Computer-Assisted Survey Information Collection," Working Paper 19. Washington, DC: Statistical Policy Office, Office of Management and Budget, May, 1990.

2. Thornberry, Owen T., Jr, An Experimental Comparison of Telephone and Personal Health Interview Surveys: Vital and Health Statistics 2(106) Government Printing Office: Washington, DC, 1987.

3. National Center for Health Statistics and Bureau of the Census, REPORT OF THE 1987 AUTOMATED NATIONAL HEALTH INTERVIEW SURVEY FEASIBILITY STUDY; An Investigation of Computer Assisted Personal Interviewing. NCHS Working Paper Series, Number 32. Hyattsville, MD: NCHS; November, 1988.

4. Gardenier, J. S., Kovar, M. G., Shumate, P., and Adams, C., CONCEPT DESIGN FOR THE NCHS COMPUTER ASSISTED PERSONAL INTERVIEWING (CAPI) SOFTWARE SYSTEM, Working Paper 34. Hyattsville, MD: National Center for Health Statistics, September, 1990.

5. Gardenier, J. S., Biggar, R. S., Adams, C., and Shumate, P., SOFTWARE REQUIREMENTS SPECIFICATIONS AND DESIGN OVERVIEW: COMPUTER/LAPTOP ASSISTED SURVEY INFORMATION COLLECTION (CLASIC) INTERVIEWERS' SUBSYSTEM, Working Paper 36. Hyattsville, MD: National Center for Health Statistics, January, 1991.