Approaches for Incorporating User-Centered Design into CAI Development

Bill Mockovak & Jean Fox
Behavioral Science Research Center
Office of Survey Methods Research
Bureau of Labor Statistics

Abstract

Almost all survey interviews are now conducted using some type of computer-assisted interviewing (CAI) software. However, unless CAI instrument design and usability are considered, even the most carefully worded questions can yield questionable data. Ample evidence exists that usability is enhanced when user needs are considered early and continuously throughout the software development process. In the development of complex computer-assisted interviewing instruments, this means bringing interviewers into the development process as soon as possible. But developing complex CAI applications poses special demands, because interviewers are often highly diverse in computer skills and geographically scattered, which makes obtaining their input more challenging. This paper discusses different approaches that have been used to address instrument design and to incorporate user-centered design principles into the development of complex computer assisted personal interviewing (CAPI) instruments. Examples from the Consumer Expenditure Quarterly Interview and the Commodity & Services Pricing survey will be cited. Besides describing possible approaches that could be used to encourage user-centered design, this paper will also present evaluation instruments and methods that have been used to quantify the success of usability-design efforts.

Introduction

The routine use of computer-assisted interviewing (CAI) has led to increasingly complex data-collection instruments, as questionnaire designers have learned to take advantage of rapidly changing technology. Concomitantly, this trend has led to a variety of new challenges for survey methodologists ranging from the programming, testing, and debugging of data-collection instruments to the design of user interfaces for interviewers and respondents (for self-completed questionnaires).

In the course of moving surveys to CAI, some survey sponsors and managers have been surprised to learn that the process of developing, testing, and implementing complex instruments[1] is far more difficult, time consuming, labor intensive, and expensive than anticipated. These managers had hoped that eliminating paper surveys would save time and money, but they failed to recognize the necessary up-front development costs required to achieve these savings. As a result, survey managers and methodologists have looked to other fields of expertise to help resolve these issues in a more timely and cost effective manner. For example, there have been some attempts to draw more heavily on the expertise available in the field of computer science for guidance in addressing problems associated with instrument design, development, and testing.[2] Similarly, survey designers have turned to the field of human-computer interaction for direction on constructing user interfaces (Schneiderman, 1992). Although instrument development and testing are critically important tasks, the focus of this paper is on one aspect of the development process that often receives less attention: the design of the user interface and its associated functionality.

Some Historical Context

To provide some historical context, in a DOS environment, screen design and user functionality in CAI were highly constrained by the software package used. In fact, an apt analogy might be that moving from DOS to Windows-based CAI is comparable, in terms of increased functionality and capability, to the transition from typewriters to word processors.

Although there was not much flexibility in DOS to begin with, what little existed was oftentimes sacrificed in the interests of simplifying programming and trying to get a workable instrument into the field as soon as possible. On the positive side, the inherent inflexibility of DOS-based instruments simplified decisions about screen design and functionality (simply because many options were not available). But long and complex questionnaires still posed significant usability challenges to interviewers, especially when the interviewer wanted to move among sections to change answers or to back up to check on previous entries (these types of actions are generally referred to as interviewer "navigation").

---

[1] An "instrument" refers to the automated survey questionnaire.

[2] Committee on National Statistics, The National Academies. "Workshop on Survey Automation," April 15-16, 2002, Washington, D.C.

In addition to serious problems with navigation, data entry was often limited in DOS instruments to one data-entry item (survey question) per screen.[3] For short, simple questionnaires, this was not a problem, but for longer, complex questionnaires, the one question-per-screen approach led to a severe segmentation of the questionnaire content. It also made it exceedingly difficult for interviewers to familiarize themselves with the content or to develop a conceptual map of the overall structure of the questionnaire. Due to such limitations (and associated problems with software), there was a general reluctance among interviewers to attempt even simple actions such as backing up a few items either to check or to change a previous answer. In fact, to avoid potential problems and confusion, some survey training actually directed the interviewers not to back up. Some complex CAI instruments went even further. They actually prohibited backing up once a module or section had been completed.

Other complaints associated with the design limitations of DOS instruments also surfaced. A common one was generally referred to as "interviewer dependency," which occurred when interviewers would blindly ask questions even if they were totally inappropriate for a respondent (this would usually happen after a key-entry error put the interviewer on the wrong path in the instrument). In these situations, survey observers complained that it seemed like interviewers had stopped thinking and were robotically reading the survey questions.

Considering these limitations, DOS-based instruments for complex questionnaires were not tools that enhanced an interviewer's ability to do the job. Instead, they restricted flexibility, sometimes added to the length of the interview (Fuchs et al., 2000) and made the interviewing task more difficult. As Woods has noted (2002), "… in design, we either hobble or support people's natural ability to express forms of expertise." In addition to worrying about obtaining quality data from respondents, interviewers now had the additional burden of manipulating a cumbersome data-collection instrument and hoping that the computer would not malfunction before the end of the interview (for example, due to battery failure).

Despite the potential problems, interviewers still reacted positively to the introduction of computers into their work (Couper and Burt, 1993). However, their positive reactions did not mean that they were necessarily happy or satisfied with CAI as a data collection tool. Instead, feedback from interviewers clearly indicated that they were willing to live with current shortcomings if they thought survey managers were actively working to address their concerns and to improve known deficiencies.

Factors that Affect Complexity from an Interviewer's Perspective

What makes a questionnaire inherently complex? From an interviewer's perspective, the following factors are important:

- Survey content.

[3] In some CAI packages, for example, CASES 4.3, it was possible to display and enter answers to multiple items on a single screen in DOS.

- Large numbers of potential questions (possibly hundreds) so that it takes many interviews to become familiar with question content and sequence. Or, question-asking sequences that lead to infrequently collected data.
- The presence of multiple sections or modules (leading to increased difficulty navigating).
- The requirement to tailor questions for different situations "on the fly."
- Complex, difficult, or lengthy data-entry tasks.
- The use of rostered questions, or question-asking sequences that involve making a selection on a screen, leaving that screen to ask additional questions about the selection, then returning to that screen.
- Variations (inconsistencies) in screen design, data entry, or user functionality resulting from the use of multiple programmers or from a failure to adhere to established standards and conventions.
- The use of tables, grids, or screens that scroll vertically or horizontally.
- An overuse of edits that require interviewer intervention.
- Multiple ways exist to do something (for example, navigate, enter missing data, etc.)
- Sub-sampling of respondents.
- Presence of dependent data.
- Ability to "spawn" cases (create new cases on the fly)

Ironically, as noted previously, questionnaire complexity has been purposely increased at times to simplify the demands of instrument programming. Although this simplifies the design and programming of the data-collection instrument, it does so at the expense of the interviewer, who then has to deal with respondents and their resulting frustration. Moreover, besides placing increased demands on the interviewer, these design decisions often result in approaches that ask the respondent to provide data in a rigid, non-conversational, manner.

An example of this type of tradeoff occurs when deciding how to obtain demographic information. For example, using two different instrument designs (programming approaches), demographic questions could be asked either on a person-by-person basis or a topic-based approach. For the "person-based" approach, something like the following line of questioning would result for two people named Robert and Suzanne. For Robert, the question sequence would be, What is Robert's age? What is Robert's race? What is the highest grade of school Robert has completed? etc., for all demographic items. For Suzanne, the identical sequence would result, What is Suzanne's age? What is Suzanne's race? What is the highest grade of school Suzanne has completed? etc., and so on, for each member of the household.

As an alternative, these questions could be asked using a "topic-based" approach that might ask the demographic questions as follows: What are the ages of the people in your family? What is the race of each person? What is the highest grade of school each has completed? Or, a slight variant would be "Has John ever been divorced "How about Mary?" "And how about Tom?" etc. This version is clearly more conversational than the person-based approach.

With a person-based design, unless interviewers can retain the information in memory, the respondent cannot report information for more than one person at a time, something that could be done easily on a paper questionnaire. In fact, on a paper questionnaire, the interviewer could use either, or both, approaches. Because some monitoring approaches (e.g., in CATI centers) require interviewers to ask each question as it appears on the computer, a very rigid, repetitious, non-conversational interviewing style results for the person-based approach. Can such a simple change have an impact? Some research suggests that a topic-based approach can increase interview efficiency, is preferred by both interviewers and respondents, reduces unit nonresponse, and, with income items as an exception, reduces item nonresponse (Moore and Loomis, 2001).

In addition to general instrument characteristics affecting complexity, other types of software or design deficiencies might also cause problems or confusion. A partial list includes:

- Lack of robustness in data entry. As an example, holding down the Enter key too long (or some other key) might result in unintentional entries for more than one question/item and move the interviewer to an unexpected question.
- Unexpected changes in screen design or layout. These types of changes cause the interviewer to "reorient" to assess the situation in order to determine what to do next, thereby adding a delay to the interview.
- Inconsistent precodes. For example, 88, 888, or 888 could all mean a "don't know" entry, depending on the question.
- Use of confusing error or edit messages. The default edit messages in some CAI packages are so confusing that interviewers routinely ask for supervisor assistance when they see them.[4]
- Inconsistent navigation functions. For example, pressing the Page Up key might take you to the first item on one screen, but to the middle of the items on a similar screen.
- Overly complex, confusing, or superfluous interviewer instructions.
- Instrument bugs or idiosyncrasies. Rather than fix 'everything,' problems ranging from minor typos to data-entry 'work-arounds' may be left in an instrument and addressed in training.

Importance of Usability

As an increasing body of research has clearly demonstrated, usability affects ease of learning, efficiency of use (including error frequency and severity), memorability (how easy it is to use software after a period of disuse), and subjective satisfaction.

Nonetheless, in many CAI development efforts, problems that are deemed minor or inconsequential are sometimes not fixed or left for "future revisions." Unfortunately, although these types of minor fixes might appear inconsequential and not be considered "show stoppers," the accumulation of these problems can affect an interviewer's attitude

[4] Based on an observation of a 2002 interviewer usability test for the American Time Use Survey, which used Blaise 4 Windows default pop-up error messages (soft and hard edits) in the instrument.

toward the instrument and introduce general inefficiencies into data collection. Laboratory studies have shown that enhanced screen designs (for example, the use of different fonts, highlighting, color, graphical features, special formatting, etc.) are preferred by users, and experimental results suggest that instruments designed using these enhancements are easier to use and require less time to complete (Beatty et al. 2000). Moreover, even small variations in screen design have been shown to affect the amount of time required for different actions (Gray and Boehm-Davis, 2000).

Seemingly minor variations in question design can also have significant effects on the data obtained (Frazis and Stewart, 1998). The following illustration shows a question that appeared in the Current Population Survey (CPS), which is used to produce monthly estimates of employment and unemployment for the U.S.

> Did you ever get a high school diploma by completing high school OR through a GED or other equivalent?
>
> (1) Yes, completed high school
> (2) Yes, GED or other equivalent
> (3) No

Previous questions in this instrument had routinely used 1 = yes and 2 = no as response options, so that interviewers got in the habit of entering a 1 for yes and a 2 for no.

Of those respondents asked this question, about 12 percent purportedly responded with Option 2, but by using external data sources, Frazis and Stewart concluded that almost all of these responses were due to spurious data entry by the interviewers. The question format shown in the illustration resulted in an estimate of 4.8 million additional GEDs in the U.S., when the true population estimate was closer to 400,000. Therefore, a slight change in question design, but a serious violation of a basic usability principle (maintain consistency) led to a significant impact on the resulting data.

With the switch to Windows operating systems and the advent of CAI packages and programming languages that offer the use of graphical tools, screen and instrument design have become much more flexible. These innovations have had a major impact on screen design and decisions about usability. For example, with the use of more modern operating systems, the following enhancements (this is not a complete list) are now possible with the use of graphical design elements:

- Type and size of font can be varied easily on the same screen and among screens.
- A variety of colors can be used.
- Shading and bolding can be used (bolding was also easily done in DOS).
- Menus and drop-down lists are available.
- Tabs are available for identifying sections of instruments and for navigation.

- Different types of section headings can be used.
- Icons can be easily used for representational purposes (e.g., for instructions, to indicate that help or additional information is available).
- "Windows," pop-up windows, or panes can be used to present additional information and to make more screen space (real estate) available.
- More complex data-entry formats (e.g., grids or tables) can be easily used.[5]
- More screen real estate can be made available through scrolling.
- Data entry can be done using the mouse or keyboard (or even voice).
- Multimedia applications are readily available.

However, with this additional flexibility has come increased responsibility for a variety of decisions concerning basic screen design and interviewer functionality. Some organizations have already codified their screen standards for "Windows-based" instruments.[6] Nonetheless, despite the recent advances made in well-known CAI packages such as Blaise and CASES, some survey sponsors have opted not to use these packages because they still consider them too restrictive and inflexible in meeting the demands of their data collectors. For example, two separate CAI data collection efforts for the Bureau of Labor Statistics' Consumer Price Index use graphical interfaces designed using Visual Basic.[7]

Implementing User-Centered Design for Complex Instruments

Usability can be assessed using a variety of techniques, but there are three general approaches (Armstrong et al., 2002):

1. Surveying techniques - including the use of questionnaires, interviews, and direct observation.
2. Inspection techniques - including standardized reviews, comparing prototypes, and heuristic evaluations.
3. Testing techniques - including modeling and simulation, think-out-loud, and experimental testing.

The key user in most computer-assisted interviewing is the interviewer (self-administration of surveys poses other, unique challenges). Therefore, the long-range goal should be the development of a data-collection tool that makes the interviewer's job easier, not harder. Unfortunately, as noted previously, instrument development efforts have often failed to achieve this goal.

To ensure that an instrument helps, rather than hinders an interviewer, a process called *user-centered* design should be implemented to parallel the instrument development process for complex instruments.

---

[5] Some DOS packages allow the use of grids but with limitations.

[6] "Specifications for Authoring CE/Blaise Standards." Internal Census Bureau document, February 10, 2000.

[7] The Commodities & Services Survey and the Housing Survey instruments.

What Are the Prerequisites and Basic Steps for Implementing User-Centered Design?

The user-centered design process assumes the use of a team process for instrument development and the presence of at least one team member who is familiar with user-interface design. Although, ideally, this person would have received special training in the usability field, other professionals with sufficient training can often fill the role. The user-centered-design (UCD) process consists of the following basic steps:

1. *Collect Data.* About users, their tasks, and their current work processes to determine interviewer requirements.
2. *Analyze the data.* To determine how to design the data collection instrument to meet the interviewer usability requirements. Identify key functionality that is required.
3. *Design and Develop.* Develop prototypes of the data-collection instrument, or separate modules, and obtain early feedback from the interviewers. Some difficult sections might require the development of multiple prototypes. To keep costs low, one can start with low fidelity paper prototypes, so you can get feedback before you make significant investments in programming.
4. *Test & Obtain Feedback.* Develop testing scenarios based on data collected in Step 1. Design and conduct iterations of usability assessments and tests.
5. *Follow-up.* Conduct follow-up studies that measure usability and user satisfaction.

In the development of survey instruments, the keys to the success of this approach are the early involvement of interviewers in the design process, mechanisms that allow for continual feedback over the development cycle of the instrument, and the active involvement of the instrument programmer (author) in the process.

Collecting and Analyzing the Data

A variety of methods exist for collecting data about user requirements. If a paper questionnaire or computer-assisted interview already exists, the data collectors will be a good source of suggestions for critical requirements and potential pitfalls. Other methods, such as direct observation of data collection and personal or group interviews, can also be used to provide supplementary information.

When obtaining this information, it is important to try to obtain input from a representative sample of interviewers, especially interviewers who vary in job and computer experience. Also, any recommendations made by interviewers should be reviewed carefully. There are likely to be cases where interviewers provide conflicting recommendations or where they have no experience upon which to base their suggestions.

To illustrate why blanket acceptance of all requests should not be the rule, in one project interviewers who had been using paper questionnaires clearly indicated, when asked, that they wanted only one survey question displayed on a screen and as much white space as

possible. Based on this input, an early CAI system was developed that met this basic requirement only to discover that as the interviewers gained more experience with the data-collection tool, they soon asked that as many questions as possible be placed on one screen. Unfortunately, due to time constraints and dwindling resources, this DOS-based system could not be modified to accomplish this goal. With hindsight, it should have been clear that interviewers were being asked to give input about something that differing amounts of experience would affect. Therefore, steps should have been taken to obtain additional feedback after the interviewers had gained more experience and before software functionality had been locked in.

Use of Prototypes

As noted previously, complex instruments are frequently divided into sections or modules, but these modules often differ significantly in their design and usability challenges. In older CAI instruments, it was not uncommon to develop the entire instrument and then begin an extensive testing and debugging process. However, with more modern CAI packages, instrument sections can be developed and tested separately and, if necessary, multiple iterations of prototypes can be developed (although in modular instruments, integration and testing of the complete instrument remains a critical step).

Why is iterative development and testing so important in the development process?

- A large, possibly unwieldy project can be broken up into steps that are more manageable.
- Critical design problems can be identified early and, therefore, addressed early in the development process.
- Changes are more likely to be made if adequate time is built into the schedule for proper development and testing of instrument modules. As time and production pressures build, significant change becomes less likely.
- Design approaches that work in one section or module can be used in others (i.e., code sharing). Similarly, design approaches that do **not** work are not used throughout the instrument.
- If desired functionality cannot be attained, there is time to develop workable alternatives that do not compromise data quality.
- Identifying problems early and fixing them means less rework and retesting later. Also, it is much cheaper to fix problems prior to production, rather than trying to fix problems once an instrument has gone into production (CNSTAT, 2002).
- Testing is generally easier for modules.

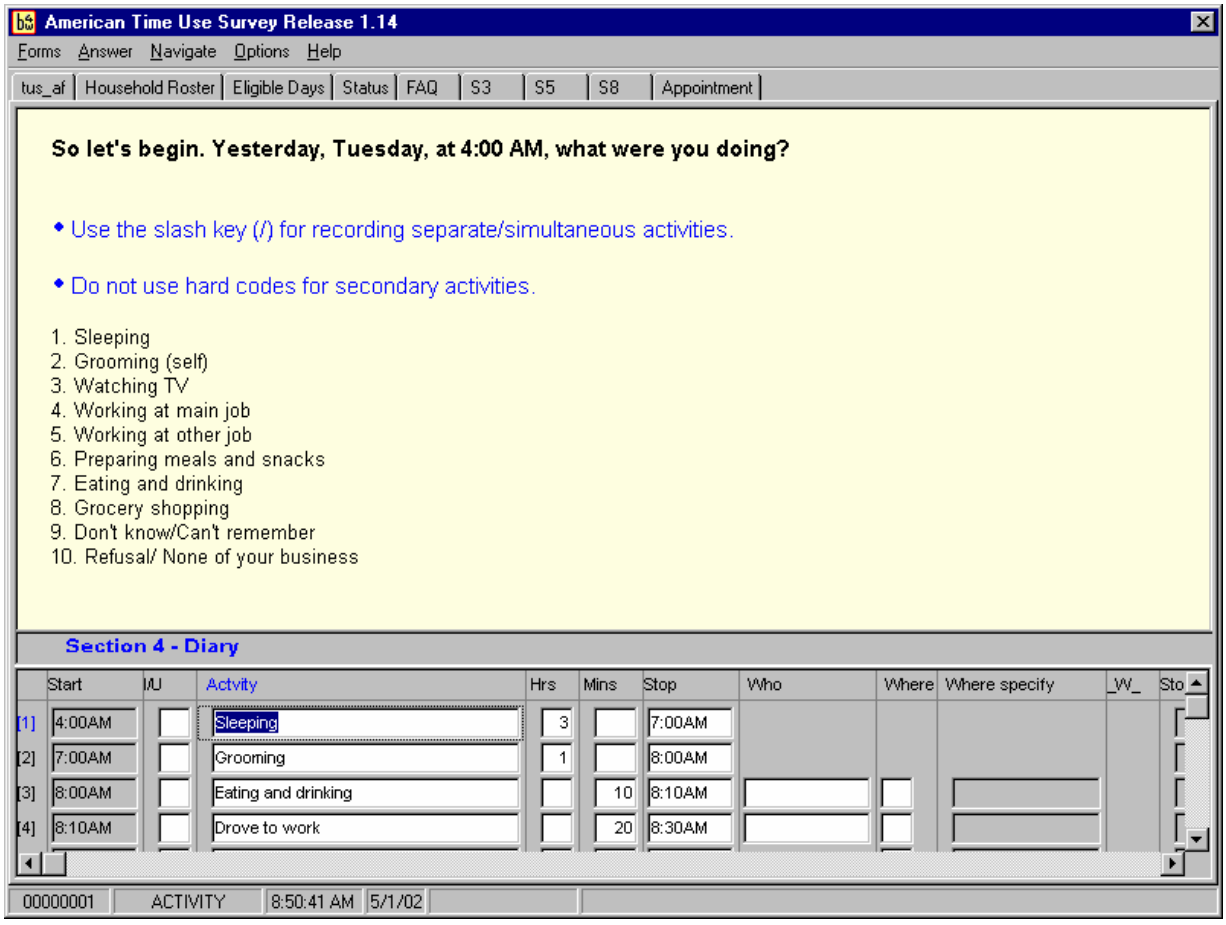
Prototype of a Diary Data-Collection Screen

The illustration that follows shows an early prototype[8] developed for one key section of the American Time Use Survey. In this survey, respondents are asked to report their

[8] This is a Blaise 4 Windows prototype.

activities for a 24-hour period: from 4 a.m. the preceding day to 4 a.m. on the day of the interview.

The prototype illustrates some of the graphical features that can be easily used in Windows-based instruments. These include the use of tabs to identify different sections, menus for added functionality, shading to designate *read-only* information (for example, the Start times for activities), the use of a grid or table format that allows multiple questions and entries on a single screen, horizontal and vertical scrolling of the table, very flexible data entry (a duration of an activity could be entered by typing a value in the hours column, the minutes column, both the hours and minutes columns, or the Stop column), special section headings, and data entry can be accomplished using either a keyboard or a mouse. [9]



Once a prototype has been developed, interviewer feedback can be obtained using one of three basic approaches:

[9]Colors were used but may not show up depending on the method of reproduction of this paper.

1. The interviewers can be brought to a central location.
2. Usability investigators can travel to the interviewers.
3. Software can be sent to the interviewers to evaluate.

The benefits of conducting usability tests are varied. If done correctly, including users in the design process improves communication and leads to a better buy-in from interviewers for the final product. Therefore, even if all desired functionality cannot be provided, reasons can be given to the interviewers explaining why, and as desired functionality is included or changes made based on their input, interviewer morale will benefit. In addition to the benefits for interviewers, usability tests also provide the instrument designers or programmers (authors) an opportunity to see their instrument in use. Therefore, summary reports of the usability test will have more impact if the developers were able to observe some of the reported difficulties.

Example of Usability Testing in Survey Instrument Development

Regular usability tests have been part of the development effort for several BLS instruments. The procedures used and the benefits gained will be discussed next for two BLS surveys: the Consumer Expenditure Quarterly Interview survey (CEIS), and the Commodities and Services (C&S) survey, both of which are done for the Consumer Price Index (CPI).

CEIS Instrument Development

In the CEIS, three separate usability tests were held about 6 months apart and involved bringing in 12 Field Representatives (interviewers), one from each of the Census Bureau's 12 regional offices, for several days at a time.[10] In addition, a pretest (300 cases) and a dress rehearsal (3,000 cases) were also conducted and presented two more opportunities for obtaining usability data and interviewer reactions.

Because a completely new CAI package was being used (Blaise 4 Windows), the initial test included the instrument's simplest sections and focused on obtaining interviewer feedback on decisions made about screen design and layout. Test 2 included the most difficult sections of the instrument, and focused more heavily on data entry and navigation issues. Test 3 included all sections of the instrument, plus a rudimentary case management system. At each testing step, the instrument included all of the sections (modules) from the previous test, plus several new sections and enhanced or new functionality added in response to interviewer feedback obtained in the preceding test. For interested readers, Attachment 3 presents suggested steps that could be followed to conduct a usability test of a survey computer-assisted interviewing instrument.

Although the primary purpose of the in-house tests was usability testing, they involved so much hands-on testing that instrument problems (bugs) were also inevitably uncovered, so extensive testing was an important side benefit. The in-house tests typically lasted

[10] See Reed (2000) for a detailed description of the usability test protocol, evaluation instruments, and results. The Census Bureau collects the data for this BLS survey.

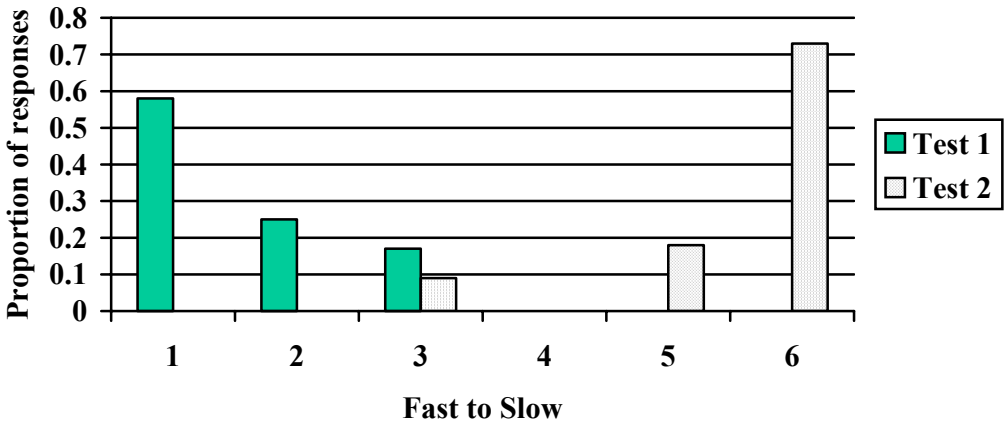
anywhere from 3 to 5 days and involved extensive interviewer interaction with the instrument that was controlled through the use of scripted interviews or fact sheets, although time was also allowed for significant amounts of unstructured testing and exploration of the instrument.

As an aside, this extensive testing effort would probably not be required for most surveys, even if long and complex instruments were involved. However, at the start of this project, several survey experts warned that they thought the Consumer Expenditure Quarterly Interview would be impossible to do on a computer, and an initial attempt to convert the survey using a DOS-based CAI package confirmed this skepticism as insurmountable design, programming, and usability problems were encountered. This experience, along with the desire to gain user acceptance of the CAPI instrument, led to the decision to implement an extensive usability testing effort.

To obtain comments from Field Representatives (FRs), two documents – "Independent Testing Reports" and "Evaluation Forms" – were used in each of the in-house tests. The independent testing forms were simple open-ended reports that enabled the Field Reps to report problems with screen design, usability, and instrument problems as they encountered them while working through scripts or unstructured testing. These forms were summarized after each test, and usability problems identified (testing errors were also summarized). The evaluation forms were used in both the in-house and field tests and varied depending on the objectives of the test. They are described next.

As part of each of the tests, each participant completed a summary evaluation form. In addition to a variety of Likert-type attitude items and open-ended questions for obtaining feedback about training, the general condition of the instrument, and a variety of other issues, a separate usability rating scale (see Attachment 1) was also completed and kept fairly consistent among the tests. Analysis of the usability-rating scale led to a ranking of usability features (see Attachment 2), which could then be compared with other evaluative feedback obtained using other means (e.g., group debriefing, individual observations, etc.). Therefore, a variety of methods were used to obtain feedback, and a qualitative assessment told us if we were hearing the same problems. However, an advantage of using the usability score rankings was that they were easy to interpret, some changed dramatically between tests, and these results could be shown to survey program managers to clarify problems and to buttress the need for additional changes to the instrument.

As an example of the type of feedback that could be provided, the most serious problem in Test 1 involved the ease with which errors could be fixed (see Attachment 2). On the other hand, instrument speed was rated as one of the better features in this same test. However, in later tests, as additional (and more complex) modules were added and the size and complexity of the instrument increased dramatically, the speed of the instrument became the foremost usability concern (although "fixing errors" remained a generally difficult task in all tests). The following chart shows more dramatically how ratings



about the speed of the instrument changed between Tests 1 and 2, where a higher scale rating indicates a problem.[11]

Looking at standard deviations of individual ratings on the scale can also provide some useful information. Small standard deviations indicate greater consistency in participant reactions, and along with positive ratings, suggest that a feature is not a problem, whereas small standard deviations and negative ratings indicate the opposite.

As a consequence of these findings, priorities for instrument enhancements were made clear, and data like these played an important part in convincing survey managers of the urgency and importance of the desired changes. In addition, the scale ratings provided evidence that overall usability was acceptable. However, since the same group of interviewers had participated in all three usability tests and the potential for a Hawthorne-type effect [12] existed, the usability rating scales were used again in a follow-up field test and a dress rehearsal with additional interviewers who had not been part of the in-house testing process.

As shown in the *Field Test Mean* column of the table in Attachment 2, some of the key usability problems (e.g., speed of instrument and fixing errors) that had been identified in the in-house tests also topped the list of problems in the field test. But the results also suggested that navigation was more of a problem in a production setting. This same pattern of results occurred in the Dress Rehearsal.

It is interesting to note that on almost all comparable items, the usability ratings were lower in the field tests than in the in-house tests. Less positive ratings in the field tests may have resulted from the greater demands of production interviewing, but also from the inclusion of interviewers (and supervisors) who had not been given specialized attention over a relatively long period of time. Also, the higher ratings from the in-house

[11] An independent samples t-test indicated the mean ratings differed significantly; t= -13.916, p<.000.
[12] The specialized attention given to interviewers over a relatively long period of time could be expected to influence their reports and ratings.

test participants could have resulted from their extended participation, which led to a feeling of ownership of the design.

Psychometric Properties of the Usability Rating Scale

As far as the psychometric properties of the usability rating scale are concerned, an internal reliability coefficient (coefficient alpha) of 0.91 was computed in Test 2 (N = 12) and a value of 0.89 (standardized alpha = 0.90, N = 119) was computed for the Dress Rehearsal. For its proposed use, this meets the minimum standards proposed by Nunnally (1967, Page 226).

Although the number of cases was limited in the Dress Rehearsal, an exploratory factor analysis was done to investigate the internal structure of the scale. At first, a variety of trial methods (principal components, principal axis, etc.) were used to determine the optimal number of underlying factors. The final solution used principal axis factoring and direct oblimin rotation (for this solution, missing values were replaced with mean values, resulting in N = 147), because it resulted in maximum interpretability of the internal structure of the scale.

Four factors accounted for about 67 percent of the total variability in the ratings. The pattern matrix is shown in Attachment 4. Based on these results, the four factors were named as follows:

Factor 1	General usability
Factor 2	Instrument speed
Factor 3	Screen clarity/usability
Factor 4	Entering and correcting data

Other Possible Measures for Assessing Usability

Another potentially useful tool for uncovering usability problems are audit or trace files produced as a by-product of computer assisted data collection. These types of files record interviewer behaviors (actions) during an interview. By reviewing these files, patterns of behavior begin to emerge. For example, Hansen and Marvin (2001) reviewed Blaise audit trails and discovered the following (this type of focused feedback led to a redesign of certain sections).

- Selected pretest items had high exit rates (interviewers left the instrument from that particular question) or a large number of abnormal terminations.
- Certain items generated high error counts. One block of items alone (asking for birth control history) accounted for approximately 29 percent of 4,525 consistency checks across all items in all interviews.
- A few interviewers had much higher or lower mean counts of consistency checks (when error message appears) per case.
- Some items took much longer than the average or expected duration.

- Some interviewers were very efficient in completing the interview, whereas others took much longer than expected.

Commodities & Services (C&S) Development

The development of the C&S data-collection instrument also took a user-centered approach. We considered the needs and skills of the users in designing the interface. We took an iterative design approach, so we built several rounds of user and system testing into the development plan.

During user testing, resource constraints prohibited bringing users into a central location. Because users worked in remote locations, an alternative approach for obtaining usability feedback had to be found. Conventional wisdom in the usability field argues that you must observe the users to truly know where the problems are. Users tend to blame themselves for problems and minimize their reporting of problems. Unfortunately, we lacked resources to allow participants to travel to Washington, D.C. or, alternatively, to have user-interface designers travel to the participants. Therefore, we explored options for making observations remotely.

The literature describes methods such as the following for observing remote users:

- Providing a button for users to click (to store video clips) whenever they experience a critical incident as they use a software prototype (Hartson, Castillo, Kelso, and Neale, 1996).
- Using a shared windowing tool and the telephone to conduct sessions remotely (Hammontree, Weiler, and Nayak, 1994).

Although we had hoped to conduct some kind of direct observations, none of these methods were practical for us. Consequently, we decided to use specially designed questionnaires[13] to solicit feedback from the users.

First, we provided users with a computer loaded with the instrument. We gave them instructions for using the instrument, conducting the test, and completing the questionnaire. Then, users collected data in the field, then returned home to complete the questionnaire. Approximately 10 people participated in each stage of testing, with more involved as we came closer to deployment.

The questionnaire was carefully constructed to solicit user feedback. The questions were very specific, so users would report on key issues (Fox, 2000; Fox, 2001). Each screen in the instrument had a corresponding page in the questionnaire. At the top of the page, a picture of the screen appeared. Below that, several questions asking the users for ratings about different aspects of the screen. At the bottom of the page, several open-ended questions allowed users to add any additional thoughts they had.

[13] See Charlton, 2002 for guidelines on developing such questionnaires.

The questionnaire was quite successful at uncovering usability problems. The ratings identified general areas of concern, and the open-ended questions allowed users to identify specific problems or to suggest alternatives. Several factors may have influenced the success of the questionnaire:

- The participants were extremely motivated to respond. They knew that they would have to use the instrument on an almost daily basis, and they wanted to be sure it would work for them.
- The specific questions and the accompanying screen shots helped participants focus on the issues.
- After the first iteration, users could see that we really listened to them, so they continued to provide input.

Thus, the C&S development project used the resources available to conduct a usability evaluation of the instrument. The questionnaire we used took some time to put together, but did not require the extensive travel or hardware costs associated with other methods of remote testing.

Other Considerations in Implementing User-Centered Design

It is worth emphasizing that the success of user-centered design rests on the successful implementation of a team process, which means that communication and organizational hurdles must be faced. For most CAI applications, the following roles are usually represented on teams:

- Survey manager
- Survey subject-matter specialist
- Usability expert or equivalent
- Survey methodologists: including the person who prepares instrument specifications
- Field manager (to represent data collection staff)
- Authors/programmers

There are numerous resources available to help structure and guide the team process (Quick, 1992; Katzenbach and Smith, 1993; Koehler and Pankowski, 1996; USDE, 1997). However, as noted, teams exist within organizations, and in addition to the challenges in using successful teams, there are numerous organizational obstacles that can stand in the way of a successful user-centered-design process. A partial list from Anderson (2002) includes the following:

- Ignorance, misunderstanding, or a different understanding of user-centered design persists.
- Credibility of messenger or specialists is questioned.
- The process itself is questioned (it's too "light-weight," "untested," "not technical enough," etc.)

- Powerful personal preferences and biases exist – for example, users are seen as part of the problem, not the solution.
- No rewards are given for attending to user needs.
- Fragmentation of the organization and of responsibilities exists.
- There is discomfort with the design process (not getting it right the first time; with the flexibility, uncertainty, imprecision).
- There are frequent organizational and personnel changes.
- Developers are insulated or isolated.
- Users are inaccessible.
- Short development schedules are the norm.
- Staff and/or budget are inadequate.
- User experience is not perceived to be an issue.

Despite the potential hurdles, user-centered-design has been shown to work in many organizations. However, to increase the likelihood of success in developing survey instruments, active management oversight and attention is required.

As shown in this paper, even when resources do not exist for the use of more traditional methods of usability testing, other less expensive methods, such as questionnaires, can be used as a substitute. Often, very simple, inexpensive methods can provide very useful feedback. The keys to success are to make the effort to obtain user feedback, especially after major design changes, and to show users that their input has had an impact on future redesigns.

References

Anderson, Richard I. "Addressing Organizational Obstacles to, and Achieving (Greater) Business Benefits from, 'User-Centered' (or 'Human-Centered' or 'Goal-Directed' or …) Design, Ethnographic Research, Multidisciplinary Collaboration, Usability Engineering, …" (evolving position paper, last updated, May 2002), http://www.well.com/user/riander/obstacles.html

Armstrong, S. D.; Brewer, W. C.; and Steinberg, R. K. "Usability Testing," in Charlton, Samuel G. and O'Brien, Thomas G. (Editors); 2002. *Handbook of Human Factors Testing and Evaluation (2nd ed.)*. Mahwah, NJ, US: Lawrence Erlbaum Associates, Publishers. pp. 403-432.

Beatty, P.; Couper, M; Hansen, S.E.; Lamias, M; and Marvin, T., "The Effect of CAI Screen Design on User Performance: Results of an Experiment." Report submitted to the Bureau of Labor Statistics, July 2000.

Charlton, Samuel G. "Questionnaire Techniques for Test and Evaluation," in Charlton, Samuel G. and O'Brien, Thomas G. (Editors); 2002. *Handbook of Human Factors Testing and Evaluation (2nd ed.)*. Mahwah, NJ, US: Lawrence Erlbaum Associates, Publishers. pp. 225-246.

Couper, M. P. and Burt, G. "Interviewer Reactions to Computer-Assisted Personal Interviewing (CAPI)," Proceedings of the 1993 U.S. Bureau of the Census Annual Research Conference.

Committee on National Statistics, The National Academies. "Workshop on Survey Automation," April 15-16, 2002, Washington, D.C.

Dumas, J.S., and Redish, J.C. *A Practical Guide to Usability Testing*. Intellect Books, Portland, OR, 1999.

Fox, J.E. "Collecting data from data collectors: Using questionnaires to gather feedback from remote users." Companion paper to poster presented at *The Usability Professionals' Association Conference*, Asheville, NC, 2000.

Fox, J.E. "Usability methods for designing a computer-assisted data collection instrument for the CPI," Proceedings of the FCSM Conference, Washington, D.C., November 2001.

Frazis, H. and Stewart, J. "Keying Errors Caused by Unusual Keypunch Codes: Evidence from a Current Population Survey Test." Proceedings of the Section on Survey Research Methods, American Statistical Association, 1998, 131-34.

Fuchs, M.; Couper, M.; and Hansen, S.E. "Technology Effects: Do CAPI or PAPI Interviews Take Longer?" *Journal of Official Statistics, Vol. 16, No. 3, 2000, pp. 273-286.*

Gray, W. D. and Boehm-Davis, D. A. "Milliseconds Matters: An Introduction to Microstrategies and to Their Use in Describing and Predicting Interactive Behavior." *Journal of Experimental Psychology: Applied, Vol. 6, No.4, 2000, pp. 322-335.*

Hammontree, M., Weiler, P., and Nayak, N. (1994). "Methods and tools: Remote usability testing." *Interactions*, 1(3) 21-25.

Hansen, S. E. and Marvin, T. "Reporting on Item Times and Keystrokes from Blaise Audit Trails." Paper presented at the 7th International Blaise Users Conference, Washington, D.C., September 12-14, 2001.

Hartson, H.R., Castillo, J.C., Kelso, J., and Neale, W.C. (1996). "Remote evaluation: The network as an extension of the usability laboratory." Proceedings of ACM CHI 96 Conference on Human Factors in Computing Systems, v.1, 228-235.

Katzenbach, Jon R. and Smith, Douglas K. *The Wisdom of Teams -- Creating the High-Performance Organization*. Boston: Harvard Business School Press, 1993.

Koehler, Jerry W. and Pankowski, Joseph M. *Teams in Government -- A Handbook for Team-Based Organizations*. Hampton, NH: St. Lucie Press; 1996

Moore, J.C. and Loomis, L.S., "Reducing Income Nonresponse in a Topic-Based Interview." 56th Annual Conference of the American Association for Public Opinion Research, 2001.

Nielsen, J. "Why you only need to test with 5 users," Alertbox, March 19, 2000, http://www.useit.com/alertbox/20000319.html

Nunnally, J.C. *Psychometric Theory*. New York: McGraw-Hill, 1967.

Quick, Thomas L. *Successful Team Building*. New York: AMACOM, 1992.

Reed, Maria. "Documentation of CE/Blaise Test 1 Activities," Internal Census Bureau report, September 2000.

Rosson, M.B. and Carroll, J. *Usability Engineering: Scenario-Based Development of Human-Computer Interaction.* Morgan Kaufmann Publishers, October 2001.

Shneiderman, B., *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, Second Edition, Addison-Wesley Publishing Company, Reading, MA, 1992.

U.S. Department of Education. Team Handbook. Washington: Department of Education, 1997.

Virzi, R. 1992, "Refining the Test Phase of Usability Evaluation: How Many Subjects Is Enough?" *Human Factors, 34, pp. 457-468.*

Woods, David D. "Steering the Reverberations of Technology Change on Fields of Practice: Laws that Govern Cognitive Work." Plenary Address: Annual Meeting of the Cognitive Science Society, August, 2002.

Attachment 1
Sample Usability Rating Scale

Instructions: Please use the following scale to rate general features of the CAPI instrument. Enter only one "X" in each row. Note that a rating of "1" is good, and a "6" is bad.

		"Good"	**1**	**2**	**3**	**4**	**5**	**6**	**"Bad"**
1.	Knowing what question to read to the respondent.	Clear							Confusing
2.	Clarity of FR instructions (blue text)	Clear							Confusing
3.	Clarity of "pop-up" messages (soft & hard edits)	Clear							Confusing
4.	Determining valid answers for questions.	Easy							Hard
5.	Locating the cursor on each screen.	Easy							Hard
6.	Fixing errors on the same screen.	Easy							Hard
7.	Backing up to fix errors on another screen.	Easy							Hard
8.	Using the function keys.	Comfortable							Uncomfortable
9.	Using the arrow keys.	Comfortable							Uncomfortable
10.	Overall screen layout.	Clear							Confusing
11.	Overall data entry.	Easy							Hard
12.	Overall speed between screens.	Fast							Slow
13.	Overall speed between answer spaces on the same screen.	Fast							Slow
14.	Overall navigation throughout the instrument.	Easy							Hard
15.	Overall ability to use the instrument.	Comfortable							Uncomfortable
16.	Learning to use this instrument.	Easy							Hard

Attachment 2

Summary of Usability Scale Ratings for 3 In-house and 2 Field Tests

1 6

"Good" <<----------->> "Bad"

Item Rated	Means				
	Test 1	**Test 2**	**Test 3**	**Field Test**	**Dress Rehearsal**
Overall speed between screens	1.6	5.5	4.5	4.9	3.8
Fixing errors	3.2	3.6	2.7	3.7	3.7
Overall navigation throughout the instrument	2.3	3.1	2.1	3.4	3.0
Overall ability to control the instrument	2.2	3.1	2.1	2.9	*
Overall screen layout	2.1	2.4	1.9	2.8	2.1
Finding answer categories for a question	2.5	2.0	1.9	2.3	2.3
Overall use of the instrument	1.7	2.8	1.8	2.9	2.4
Overall data entry	2.0	2.6	1.7	2.7	2.4
Using function keys	2.5	2.3	1.7	2.7	2.7
Knowing what to read to respondent	2.0	2.2	1.6	2.0	1.7
Learning Windows/Blaise	1.7	2.2	1.6	2.7	*
Clarity of FR instructions	2.2	2.1	1.6	2.7	1.9
Windows/Blaise training for this test	1.7	1.9	1.6	2.8	*
Locating the cursor on each screen	2.8	2.4	1.5	2.5	2.0
Using arrow keys	1.8	1.9	1.4	2.2	2.1
Edits	*	*	*	3.5	2.8
Usefulness of help screens	*	*	*	3.5	*
Ease of use of CE instr. Compared to DOS	*	*	*	3.2	*
Use of page up & page down keys	*	*	*	2.4	*
No. of Field Reps	12	12	12	47	147
Overall Mean	2.2	2.7	2.0	2.9	2.5

* Question not asked in test.

Attachment 3: Suggested Steps in a Usability Test.

When using the first two options, a usability test can be conducted by implementing the following suggested steps (for a more complete description of usability testing, see Dumas and Redish, 1999):

1. Plan the test.
   - Determine usability goals for the test. What is expected or desired behavior? For example, in a diary interview conducted over the telephone, a desired behavior would be that the interviewer be able to enter information at a normal conversational speaking pace.
   - Determine how many and which interviewers will participate in the test. [14]
   - Determine where the test will be done and how much it will cost.
   - Determine how many tests will be required over the development cycle and how much time should pass between tests.
   - Determine what functionality will be tested in each test.
   - Develop a debriefing protocol (list of questions to be discussed, and ground rules for participation).
   - Develop evaluation/rating scales to obtain structured feedback, and to compare ratings over the development cycle.

2. Develop testing scenarios.[15]
   - Use mock interview walk-throughs conducted by the trainer to explain basic functionality and to provide basic training.
   - Use fully-scripted interviews with paired-interviews[16] to control specific data entry paths.
   - Use semi-structured interviews with paired-interviews to lead the interviewer through certain instrument sections and situations.
   - Use unstructured interviews to allow the interviewer to explore usability issues individually.
   - Test out the scenarios ahead of time to determine timing and to uncover unforeseen problems.

3. Conduct the usability test.
   - Develop the agenda.
   - Have a trained facilitator lead the test.
   - Have trained observers watch, listen, and take notes.

---

[14] Some researchers argue that 5 are sufficient for each user population (Virzi, 1992; Nielsen, 2000). Ideally, the participants should have experience in the survey, or a survey of a similar nature. More participants can be used to ensure representation from different regions of the country.

[15] See Rosson and Carroll (2001).

[16] In a paired-interview, an interviewer plays the role of interviewer. Other interviewers or participants are provided with scripts or fact sheets and play the role of respondent.

- Pair off the interviewers. Have one interviewer play the role of "interviewer," while the other member of the pair acts as a respondent.
- Reinforce the point that the software is being tested, not the interviewer.

4. Conduct a debriefing.
   - Have the participants complete evaluation forms prior to the discussion (the ensuing discussion might cause interviewers to change their impressions or to question their reactions).
   - Give the participants the opportunity to express their reactions.
   - It is important that the facilitator leads and remains neutral in words and body language.

5. Prepare a summary report.
   - List the problems that the interviewers had.
   - Sort the problems by priority and frequency.
   - Develop solutions.
   - If changes cannot be made in future prototypes, be sure to explain why to the interviewers.

Attachment 4
Pattern Matrix for Factor Analysis of Usability Scale

	Factor 1	Factor 2	Factor 3	Factor 4
Knowing what question to read to the respondent.	-.099	-.035	.826	.062
Clarity of FR instructions (blue text)	.003	-.016	.658	-.016
Clarity of "pop-up" messages (soft & hard edits)	-.009	.064	.456	.306
Determining valid answers for questions.	.076	.004	.246	.544
Locating the cursor on each screen.	.280	.069	.421	.046
Fixing errors on the same screen.	.224	.213	.088	.466
Backing up to fix errors on another screen.	.315	.160	.015	.645
Using the function keys.	.484	.093	-.026	.279
Using the arrow keys.	.414	.021	.252	.112
Overall screen layout.	.601	.064	.436	-.263
Overall data entry.	.559	.269	.338	-.293
Overall speed between screens.	-.089	.871	-.017	.005
Overall speed between answer spaces on the same screen.	-.042	.992	-.089	.008
Overall navigation throughout the instrument.	.573	.088	.068	.228
Overall ability to use the instrument.	.840	-.025	-.107	.078
Learning to use this instrument.	.800	-.084	-.069	.036