USABILITY EVALUATION OF THE NHIS CAPI INSTRUMENT

Sue Ellen Hansen and Mick P. Couper, Survey Research Center, University of Michigan,
Marek Fuchs, Catholic University of Eichstaett, Germany
Sue Ellen Hansen, Survey Research Center, 426 Thompson Street, Ann Arbor, Michigan 48106-1248

Key Words: CAI, Usability Instrument Design

The authors acknowledge the support of the National Center for Health Statistics, under Cooperative Agreement #S278-15/15 with the Survey Research Center (James M. Lepkowski, Principal Investigator). We also acknowledge financial assistance from the Research Center for Group Dynamics development of the laboratory used for the tests described in this paper, and of the Humboldt Foundation of Germany for support of Dr. Marek Fuchs.

Introduction

Designing for usability is "...the practice of designing products so that users can perform required use, operation, service, and supportive tasks with a minimum of stress and a maximum of efficiency" (Woodson, 1981, cited by Rubin, 1994, page 10). It is generally acknowledged that designers of computer systems need to pay attention to the users of their systems. Thus, there is a large body of research on human factors engineering, user-centered design, and human-computer interaction (HCI) that focuses on users and the design of computer system interfaces. However, while HCI has gained acceptance as necessary to software development and evaluation, it has had little impact on the design of computer assisted instruments until very recently (Couper, 1994; Couper, 1997; Hansen, Fuchs, and Couper, 1997).

Computer assisted interviewing (CAI) introduces design issues not addressed in the development of paper survey instruments, especially the many ways in which technology may affect interviewer and respondent interaction and resulting data quality. Such design issues are issues of usability.

Usability research emphasizes the cognitive and interactional aspects of computer use, addressing the ease or difficulty a user has interacting with hardware and software. Difficulty arises when design features conflict with a user's goals for or expectations of the system. CAI software and instruments can vary in the degree to which they are easy for interviewers and respondents to use in the performance of their role-specific tasks in the interview.

Ease of use is determined in large part by the design of the computer interface--the display of information, availability and implementation of system features and functions, and types of feedback provided following respondent and interviewer actions. Research on computer assisted interviewing has tended to neglect the impact of CAI on users, although there are exceptions (e.g., Couper, Hansen, and Sadosky, 1997; Edwards et al., 1995). The focus primarily has been on feasibility of CAI--on technology, programming, and costs, rather than on designing for ease of use (Couper, 1997; de Leeuw and Collins, 1997).

Although research suggests that interviewers are positive toward the use of CAI (e.g., Weeks, 1992), there is evidence that they sometimes have difficulty using CAI instruments and systems (e.g., Couper and Burt, 1994). There is also evidence that some mode differences reported between CAI and paper surveys can be attributed to differences in instrument design and layout (e.g., Baker, Bradburn, and Johnston, 1995; Bergman et al., 1994). This provides support for the belief that designers of CAI instruments need to go beyond traditional research on questionnaire design, which focuses on question content and the respondent's understanding of questions. They must also evaluate the *usability* of their instruments, that is, how easy it is for users to interact with CAI instruments and systems. Interviewers and respondents are the key users of interviewer- and self-administered surveys, and survey designers should address their needs. To the extent that an automated instrument facilitates interviewer and respondent performance, data quality improvements may result.

Many of the techniques available for pretesting paper-and-pencil questionnaires, such as cognitive interviewing, can be used to evaluate CAI systems and survey instruments, including the effectiveness of CAI screen layout and design. However, while most methods for evaluating survey instruments focus on the respondent's understanding of the questions, usability evaluation focuses on the interviewer's interaction with the CAI system and survey instrument. This shifts the focus of CAI research from the respondent to the interviewer and from system feasibility and functionality to design of instruments from the interviewer's perspective. In self-administered surveys, this shift in focus is from the respondent as information processor to the respondent as both information processor and CAI system user. This view acknowledges that the interviewer plays an important role in mediating between what the designer has embedded in the instrument and the respondent, and that the instrument and computer affect interaction between the respondent and interviewer.

One method used in HCI research to evaluate usability is the laboratory based usability test. In such tests, people are observed in a controlled setting as they

use computer systems. There are less costly methods of usability evaluation (Nielsen and Mack, 1994). However, since usability testing is the only method that involves the users themselves, it is particularly effective at identifying serious and recurring usability problems. As Nielsen and Mack observed, "One cannot expect ... to address all usability issues when the evaluators have no knowledge of the actual users and their tasks" (1994:45). This appears to be a common problem in CAI design, where programmers are more removed from the users of their instruments than were paper-and-pencil questionnaire designers. Thus, it is valuable to include at least a small usability-testing component in studies that heavily rely on other instrument evaluation techniques.

This paper describes the results of a study in which usability tests were used to evaluate the instrument for a computer assisted personal interview (CAPI). Observations, coding of interviewer and respondent behaviors during the tests, and an analysis of CAPI screen characteristics were used to identify design problems.

Data and Analyses

Data examined in this paper were collected as part of an evaluation conducted for the National Center of Health Statistics (NCHS) of the 1997 CAPI instrument of the National Health Interview Survey (NHIS). The NHIS is an ongoing survey of health-related issues in the United States, and has been conducted continuously since 1957. The U.S. Bureau of the Census is the data collection organization for the NHIS, under contract to the NCHS. The survey has undergone a period of redesign over the past few years as it converted the paper-and-pencil personal interviewing (PAPI) instrument to a CAPI instrument. The redesign also involved extensive alterations to the questionnaire and changes to the sample design. Phase I of the redesign, during the first six months of 1996, involved 16 interviewers using the CAPI instrument. In Phase II, during the last half of 1996, all NHIS interviewers conducted about half of their sample cases using CAPI, and the remainder using PAPI. Finally, in Phase III, beginning in January 1997, the entire NHIS sample was switched to CAPI. The focus of this paper is on the Phase III CAPI instrument. The NHIS CAPI instrument is programmed in CASES version 4.2.

As part of the evaluation of the Phase III NHIS CAPI instrument, Detroit area U.S. Bureau of the Census interviewers conducted 38 NHIS CAPI interviews in a laboratory setting (Hansen, Fuchs, and Couper, 1997). Observations from these interviews revealed a number of usability problems, that is, difficulties interviewers experienced while trying to perform the tasks required of them, such as reading questions, following instructions, and using CAI functions.

Further analyses were conducted to provide quantitative data about interaction between the interviewer and respondent in the usability interviews, and to identify more concretely the types of problems revealed through CAI instrument usability testing. The primary source of data for these analyses was a videotape of scan-converted images of the laptop computer screens as they appeared to the interviewer during each interview. Two additional videotapes, one of the interaction between the interviewer and respondent, and one of the interviewer's hands and the computer keyboard, were also available for analysis, as were the audiotapes used for behavior coding these interviews.

Each computer screen accessed in the interview was coded to indicate the occurrence of specific interactional events that may occur in either interviewer-respondent or interview-computer interaction. Event coding goes beyond traditional behavior coding of interviews (Oksenberg et al., 1989), attempting to capture occurrences such as computer beeps, backups, and extended silences that might reveal usability problems. It also broadens definitions of behaviors in an attempt to capture other aspects of screen design beyond the format of question text, such as the success or failure of using functions.

Table 1 lists the event codes used in these analyses. Each code is included as a direct or indirect indicator of either respondent-interviewer or interviewer-computer interactional difficulty. Events coded include behaviors that may be captured in more traditional coding of behavior in interviews, such as interruptions, digressions, task- or affect-related comments, questions not read as they appear on the screen, and so on (Oksenberg et al., 1989). Such codes, designed primarily to capture question wording and response problems, focus on interviewer-respondent interaction, but they may also reveal interface design problems. Also coded are events such as problems with data entry, reference to an interviewer aid, and prolonged silence after a response is given. These codes focus more on computer-interviewer interaction, and are included specifically to identify interface design problems, but may also reveal difficulties in interviewer-respondent interaction. Laughter is included as a potential indicator of interviewer or respondent discomfort during either interviewer-computer interaction or respondent-interviewer interaction. Event coding analysis is intended to identify sequences of interaction for additional review, to determine if they reveal usability problems.

Coders viewed the videotape of the scan-converted screen images for each interview. For each screen displayed, they listened to the interviewer-respondent interaction while observing the screen displayed to the interviewer during the interaction. For events that

Table 1. Event Codes Used in Evaluation

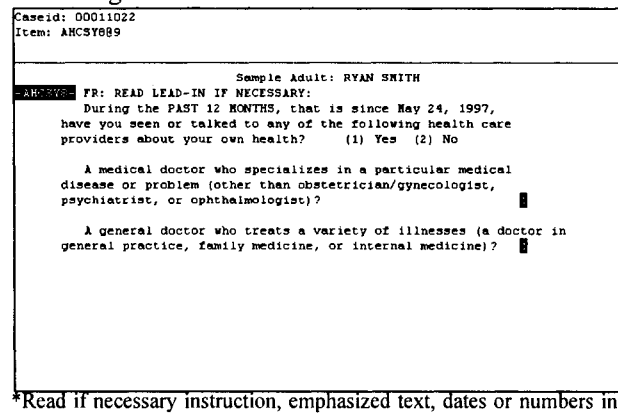
A	Reference to interviewer aid
T	Problem reading text
N	Task not complete
P	Probe for additional information
F	Feedback
C	Affect- or task-related comment
D	Any other comments or digressions
Q	Question comprehension problem
R	Response problem
I	Interruption
L	Laughter
S	Silence
E	Problem recording or entering data

occurred one or more times (for example multiple interviewer probes), they assigned the appropriate event code once (for example one "P"). Thus, the unit of analysis is each screen that is presented to the interviewer during the interview. A screen may not have any events associated with it. Screens may be coded more than once, such as screens that appear for each member of the household. Screens may also be coded more than once if an interviewer enters a response and subsequently returns a second or third time to review the question or instruction or to change a response. If the interviewer started to read a question, stopped, and then backed up to review the previous screen, the screen from which she backed up would be coded as "INE" (interruption, question reading not completed, and backup. As with traditional behavior coding, such event coding is subject to reliability problems, since it depends on coder judgment. Future analyses will assess the reliability of these event codes.

It was hypothesized that certain interactional difficulties are likely to be associated with particular screen types. For this reason, in addition to event coding, each screen was coded to indicate the features of the screen and its relative complexity. Characteristics coded included such features as text enhancements, multiple-response, multiple-item, types of instructions, number of response options, help indicators and so on. There were a total of 17 characteristics. The number of screen characteristics, which ranged from zero to 11, may be an indicator of the complexity of the screen. For example, Figure 1, shows a screen, AHCSY8, with seven of the screen characteristics—header information, read-if-necessary instruction, emphasized text, dates or numbers in text, other text characteristics (optional text and slashes), multiple items, and two response options. With a variety of things the interviewer must attend to, this screen is obviously more complex than a single-item screen with simple response options and no text enhancements, instructions, or headers.

Excluded from the analyses were questions for which the laboratory setting may have resulted in events that would have not been encountered in a more natural setting. These included questions such as one that asks about the telephone number "here," and another that asks about whether the respondent owns or rents "this" home. The final data set included 11,336 exchanges, representing 471 screens across the 38 usability interviews. For any screen accessed five or more times across in the usability interviews, we calculated a proportion of times an event occurred on a screen, and then calculated standardized scores based on the mean proportions of events.

Figure 1. Screen with Seven Characteristics*



*Read if necessary instruction, emphasized text, dates or numbers in text, other text characteristics (optional text and slashes), multiple items, and two response options.

Findings

Observation of the usability tests. Analysis began with observation of the usability tests. One of the major findings from observing NHIS interviews in the usability laboratory was that there were a number of screen layout and design features used inconsistently that could lead to problems in interaction between the interviewer and respondent. Particularly problematic was the use of capitalized text and complete names for insertion or "filling" of household member names, including that of the respondent. The usability tests revealed that certain interviewers often tended to read the full name on such screens, leading to awkward interaction. Other interviewers were more apt to tailor the text by using relationship terms or first names, leading to "bad" interviewer behavior, at least as defined in standard behavior coding (see Lepkowski et al., 1998).

Capitalized text was also used for other "variable text" and for words to be emphasized, as well as for interviewer instructions (usually but not always boldfaced and separated from question text), which appeared to confuse some interviewers. Some could be heard to pause at capitalized words, as if attempting to determine how to treat them; and some did not

emphasize text that should have been. This was particularly problematic when capitalization was used for multiple purposes in the same question, such as for emphasis, introductions to the question, interviewer instructions, capitalized response options, and so on.

The NHIS usability tests also revealed a recurring problem with the use of hand card instructions. Usually a hand card instruction appeared above the question to be asked, so that the interviewer received a cue to refer to the hand card before or as she read the question. Several questions placed the hand card instruction following the question text. This sometimes led to interviewer self-interruptions, that is, stopping in mid-question to refer to the card and rereading the question, and other times to failure to refer to the hand card altogether.

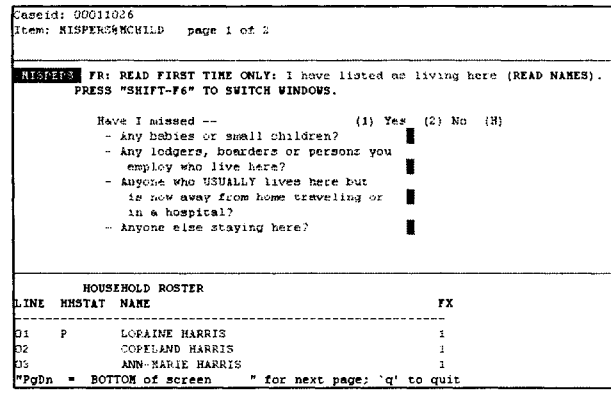
The observation of problems such as these led to a systematic review of the NHIS instrument, resulting in a number of suggestions for improved screen layout design (Hansen, Couper, and Fuchs, 1997), and NCHS has begun to implement some of the suggested changes.

One of the most glaring problems in the usability tests was a problem interviewers had using the function SHIFT-F6. This function toggles between the question text window and a roster of household members, when the full list of members cannot be displayed on the same screen (see Couper and Schlegel, 1998, for a discussion of this problem based on trace file analysis). This problem was observed in the first usability interview in which there were four or more household members, on the item MISPERS (Figure 2). The interviewer never successfully used the function, but was able to recall the full household listing from memory (see Hansen, Fuchs, and Couper, 1997, for a more complete discussion of this example).

To determine how pervasive this problem was, we reviewed the MISPERS exchange in the 18 usability interviews with four or more household members. Across the 9 interviewers who conducted these interviews, there were only five attempts, none of which was successful. Thus, not once in the usability interviews did an interviewer successfully invoke the SHIFT-F6 function.

The function SHIFT-F6 is problematic for interviewers in part because it requires the use of two function keys, making it more difficult to remember and more difficult to use. Problems might be minimized by assigning the function to a single function key. The problems with screens like MISPERS might better be solved by displaying all household information on the screen with the question, but remaining in a separate window, eliminating the need for a window toggle function. NCHS has chosen this latter solution for this screen and some others in a more recent version of the NHIS instrument.

Figure 2. Example of Screen with SHIFT-F6



Event Codes and Screen Characteristics. The event coding and screen characteristic analyses provided some quantitative support for the findings from direct observation of the usability interviews. In eighteen percent (18%) of the exchanges the interviewer failed to perform or complete a task, such as reading a question, reading a list of household members, referring to a show card, probing as explicitly instructed, probing a range, and so on. In over 27% of the exchanges, an interviewer had problems reading the text, leading to changed wording, stumbling over words, and so on. Any of the problems identified through observation could have contributed to these proportions of events.

There was silence and mumbling in nearly 14% of the exchanges, which may be attributable to a large number of checkpoints during which the interviewer is not required to interact with the respondent. Other prevalent behaviors, such as feedback (24%) and probing (11%), are not necessarily problematic. However, additional research on evaluating the usability of CAI instruments is necessary to determine whether these rates differ from rates found in paper-and-pencil interviews, and if so, to attempt to determine the factors contributing to those differences.

In order to identify the most serious problems revealed through event coding, we computed a standardized event score for each question. Questions with two standardized scores greater than or equal to 2.0 were selected (see Lepkowski et al., 1998 for a description of this technique). The 24 NHIS items or screens thus identified are listed in Table 2.

Of these items, several also could have been revealed as problematic through behavior coding, if we focus only on significant scores for respondent and interviewer behaviors such as probing, digression, question comprehension, feedback, interruption and so on. Thus items such AFLHCL, CP2ADDR, CP2NAME, CPNAME1, DPTENO, FAMINC, FWHY, HIBEV, MLTRAC may appear as problematic through behavior coding as well as usability evaluation of the same instrument.

Table 2. NHIS Screens Identified by Event Coding as Problematic (Two or More Standardized Scores ≥ 2.0)

Question / Response Only (10)	Interviewer Aid (3)	Reading Question Text (2)	Task Incomplete (5)	Laughter / Entry / Silence (6)
AFLHCL	C4	CCOLD2W	AHCNOYR	IJHOW L,E
CP2ADDR	CCOLD2W	HIBEV	CMHAGM31	IJTYPE S,E
CP2NAME	JNTYRP		HICHECK	JNTIJL E
CPNAME1			WHAT	LCASPEC E
DPTENO			WKLS	MLTRAC L
FAMINC				WHAT E
FWHY				
HCSPFYR				
HICOST				
LASTST				

Others like CCOLD2W and HIBEV, have the subject's name inserted into the text (sometimes twice), and/or other capitalized text, which may contribute to problems reading text as observed in the usability laboratory. The only screen with a high standardized score for silence was IJTYPE, which has two large open fields, which could account for extended silence while the interviewer types in text. Four high data entry error scores occurred on screens with open text fields (IJTYPE, JNTIJL, LACASPEC, and IJHOW) and one occurred on a multiple-response screen. Both types of screens could understandably exhibit higher rates of data entry problems.

Many items with two or more high standardized scores have problems that may be attributable to placement of hand card instructions. AHCNOYR, MHSAD_CK, CMHAGM21, HICHECK, and WHAT have hand card instructions that appear at the end of the question text, or embedded in other instructions. These items all exhibit high "task incompletes," which may indicate failure to refer to the hand card. There are two other screens with apparent hand card problems, C4 and JNTYRP. These screens, which did not have explicit hand card instructions, had an unusually high number of references to the hand card booklet. Both screens are preceded by introductory screens on which there is a hand card reference.

Such screen combinations may exhibit usability problems for two reasons. First, on the lead question or introductory screen, the hand card instruction often follows the text the interviewer reads, and thus may be overlooked. Second, on a followup screen such as C4, where the specific hand card is used by the respondent to provide a response, there is no indication of the hand card the respondent should be using. This makes it difficult for the interviewer to refer to the appropriate card again if necessary, whether or not she correctly referenced the card at the preceding screen. Ideally, CAI systems would make it possible to display hand card references or any other context information in the header or non-active window of the screen.

Conclusion

It appears that the event coding analysis supports the observations from usability testing of the NHIS instrument. However, it should be stressed that this was not a routine usability test, since an explicit goal has been to evaluate usability testing as a method of evaluation for survey instruments. Multiple recordings and event coding are not necessary components of a typical usability test. With a simple data-logging program used to collect very basic goal-oriented "event" data (see Rubin, 1994, for an example), the only other requirement is an observation area. Data from such a system could easily be combined with other measures of potential usability problems such as question-level time stamps and keystrokes. Thus, with an established observation area and data logging software, usability testing becomes a reasonably priced option for pretesting questionnaires.

The evidence from these analyses supports the view that usability testing identifies serious and recurring problems, such as the SHIFT-F6 problem and missing or misplaced hand card instructions. Usability testing has also been useful in detecting the reasons for recurring problems reading text, and is the only method that can provide evidence of the impact of instrument design on interviewers themselves. Many problems identified through usability test observations and event coding parallel those found through trace file analysis and behavior coding. However, some, such as the SHIFT-F6 problem, may be impossible to detect and/or diagnose with the other methods. In addition, videotaped usability tests serve as a visual record of the kinds of difficulties experienced by interviewers, providing supporting evidence sometimes necessary to convince designers and programmers that instrument changes are necessary.

The overlap with other instrument evaluation methods suggests that usability tests or interviews could supplement or comprise a portion of more typical pretest survey interviews (see Lepkowski et al., 1998, for a comparison of evaluation methods). HCI research has found that up to 80% of serious and recurring problems can be detected in as few as four or five usability tests, although 10 or more tests per treatment are recommended for experimental designs (Rubin, 1994). Thus, for example, at marginal increased cost, five out of 30 pretest interviews could be conducted in a laboratory (if available), and even audiotaped for behavior coding if behavior coding were part of the instrument evaluation.

The potential for using question characteristic data in usability evaluation has not been fully explored. There is some evidence from preliminary analyses that at least some of the characteristics examined are predictive of certain problems or events revealed in observation and event coding usability evaluation, as well as trace file analysis and behavior coding. A review of items suggests that those with a single significant standardized score were often easily explained. For example, items for which only silence appeared as a significant event most often were associated with checkpoints or open-ended questions. It is possible that problems reading text in instruments such as the NHIS are correlated with characteristics such as name or other variable text, capitalized text, and optional text. If such characteristics were found to be related to problem behaviors, such a question coding scheme may be useful in future evaluations to isolate questions for more systematic analysis and evaluation.

References

Baker, R. P., N. M. Bradburn, and R. A. Johnson (1995). Computer-Assisted Personal Interviewing: An Experimental Evaluation of Data Quality and Costs. *Journal of Official Statistics,* 10(2): 181-195.

Bergman, L. R., Kristiansson, K.-E., Olofsson, A., and Säfström, M. (1994). Decentralized CATI Versus Paper and Pencil Interviewing: Effects on the Results in the Swedish Labor Force Surveys. *Journal of Official Statistics*, 10(2): 181-195.

Couper, M. P. (1994). What Can CAI Learn from HCI? Discussion paper presented at the COPAFS Seminar on New Directions in Statistical Methodology.

Couper, M. P. (1997). The Application of Cognitive Science to Computer-assisted Interviewing. Paper presented at the CASM II Seminar, Charlottsville, Virginia, June 12, 1997.

Couper, M. P., and G. Burt. (1994). Interviewer Attitudes toward Computer-Assisted Personal Interviewing (CAPI). *Social Science Computer Review* 12(1):38-54.

Couper, M. P., S. E. Hansen, and S. Sadosky. (1997). Evaluating Interviewer Use of CAPI Technology. In L. Lyberg et al. (eds.), *Survey Measurement and Process Quality*. New York: John Wiley & Sons.

Couper, M. P., J. Horm, and J. Schlegel. (1996). The Use of Trace Files for Evaluation of Questionnaire and Instrument Design. Paper presented at the International Conference on Computer-assisted Survey Information Collection, San Antonio, December.

Couper, M.P., and J. Schlegel. (1998). Evaluating the NHIS CAPI Instrument Using Trace Files. Paper presented at the annual meeting of the American Association of Public Opinion Research, St. Louis, MO, May 1998.

de Leeuw, E., and M. Collins. (1997). Data Collection Methods and Survey Quality: An Overview. In Lyberg, L., et al. (eds.), *Survey Measurement and Process Quality*. New York: John Wiley & Sons.

Edwards, B., Sperry, S., and Schaeffer, N.C. (1995). CAPI Design for Improving Data Quality. *Proceedings of the International Conference on Survey Measurement and Process Quality, Bristol, U.K.*, pp. 168-171.

Hansen, S.E., Couper, M., and Fuchs, M. (1997). Usability Evaluation of the NHIS CAPI Instrument. Final Report, NCHS Cooperative Agreement #S278-15/15, "Preparing the National Health Interview Survey for the 21st Century."

Hansen, S.E., Fuchs, M., and Couper, M.P. (1997). CAI Instrument Usability Testing. Paper presented at the annual meeting of the American Association of Public Opinion Research, Norfolk, VA, May 1997.

Lepkowski, J.M., M.P. Couper, S.E. Hansen, W. Landers, K.A. McGonagle, and J. Schlegel. CAPI Instrument Evaluation: Behavior Coding, Trace Files, and Usability Testing. Paper presented at the annual meeting of the American Association of Public Opinion Research, St. Louis, MO, May 1998.

Nielsen, J. and R.L. Mack (1994). *Usability Inspection Methods*. New York: John Wiley & Sons, Inc.

Oksenberg, O., Cannell, C.F., and Kalton, G. (1989). New Methods for Pretesting Survey Questionnaires. Final Report, Survey Research Center, University of Michigan.

Rubin, J. (1994). *Handbook of Usability Testing*. New York: John Wiley & Sons, Inc.

Weeks, M. F. (1992). Computer-Assisted Survey Information Collection: A Review of CASIC Methods and Their Implications for Survey Questions. *Journal of Official Statistics,* 8(4): 445-465.

Woodson, W. E. (1981). Human Factors Design Handbook: Information and Guidelines for the Design of Systems, Facilities, Equipment, and Products for Human Use. New York: McGraw-Hill, 1981.