REAPING THE BENEFITS OF MODERN USABILITY EVALUATION: THE SIMON STORY

James R. Lewis
IBM Human Factors Group
P. O. Box 1328
Boca Raton, FL 33429-1328
Tel: +1 (407) 443-1066
Fax: +1 (407) 443-2778
E-mail: JIMLEWIS@VNET.IBM.COM

product usability; usability evaluation methods; personal communicators

Simon (TM-BellSouth Corp.) is a commercially available personal communicator (PC), combining features of a PDA (personal digital assistant) with a full suite of communications features. This paper describes the involvement of human factors engineering in the development of Simon, and summarizes the various approaches to usability evaluation employed during its development. Simon has received a considerable amount of praise from the industry and won several industry awards, with recognition both for its innovative engineering and its usability.

INTRODUCTION

The Simon is a cellular telephone, designed with a 36 x 115 mm touch screen (CGA resolution) replacing the standard telephone key area. Research in the usability of cellular telephones (Tsoi, 1993) has shown that many of the problems people have using cellular telephones are the result of inflexible control labeling and limited feedback. Replacing the standard key/display area with a touch screen allowed the Simon developers to create a simpler user interface for cellular telephone tasks. It also allowed the development of a suite of applications in addition to the cellular telephone, including an appointment calendar, an address book, a to-do list, a world clock, a note pad, a sketch pad, sending and receiving electronic mail, sending and receiving faxes, reception of pages, file management, a calculator, access to system settings, and security.

My first contact with the Simon development group came as a request to answer an apparently simple question: How small can a touch screen button be, and still be usable? Fortunately, I had just completed a literature review covering the results of human factors studies of touch screens from 1980 to 1992 (Lewis, 1992), so I was able to convey to Simon development that the answer to this simple question was actually somewhat complex and depended on the touch selection strategy (Sears and Schneiderman, 1989). From this start, I spent the next two years as a part of the Simon team, conducting studies and providing usability guidance. The approaches to usability engineering and assessment applied during Simon development illustrate the broad spectrum of modern usability methods, and the resulting product demonstrates the effectiveness of these modern methods. The descriptions appear in rough order of occurrence, but the activities overlapped considerably.

APPROACHES TO USABILITY ENGINEERING AND ASSESSMENT IN THE DEVELOPMENT OF SIMON

Focus Groups

After preliminary design work, an independent agency conducted several focus groups with different types of cellular telephone and computer users to help define the appropriate goals for the product.

Daily Design Meetings

Before writing any significant amount of code, the software team (including a human factors engineer and graphic designer) worked out more specific details about how to achieve the design goals. These meetings lasted for several hours every morning over a period of several months. After each meeting, the individual designers worked on their assignments, which typically involved detailed functional and task analyses. During the meetings, the designers presented their analyses and the rest of the team proposed scenarios for testing the task flows. Determination of problems with task flows in these meetings led to additional refinement of task analyses, which led to refinement of design concepts.

Literature Reviews

Literature reviews of human factors studies of touch screens (Lewis, 1992) and cellular telephone usability provided early, valuable guidance to Simon development. It is often tempting to skip the tedium inherent in a literature review, but keep in mind that it would be foolish to spend three months in the laboratory to obtain information available with an investment of three hours in a library.

Expert Evaluations of Competitive Products

Using an approach similar to Nielsen's (1992) heuristic evaluations, I conducted several expert evaluations of competitive products, both defining the sequence of steps required to perform key tasks and making note of probable problem areas. These evaluations revealed opportunities for improved design in such diverse areas as battery installation and removal, display contrast adjustment, key definition as a function of mode, setting calendar alarms, effective setting and removal of repeating meetings, and clear procedures for setting passwords and locking units.

Development of Test Scenarios

Considering the focus groups, daily design meetings, and expert evaluations of competitive products, the team developed an initial set of 38 test scenarios. By the end of iterative testing, there were 54 scenarios. As suggested by Lewis, Henry, and Mack (1990), some scenarios focused on tasks within a single application, while others evaluated work that crossed application boundaries. We used the scenarios for both gathering competitive performance and satisfaction benchmarks and for iterative problem discovery studies with development-level versions of Simon.

Competitive Usability Benchmarking

One application of the test scenarios was the determination of competitive usability benchmarks for both user performance (scenario completion times and success rates) and satisfaction. We used the After-Scenario Questionnaire (ASQ) to assess user satisfaction following each scenario, and the Post-Study System Usability Questionnaire (PSSUQ) to assess more global usability satisfaction following the completion of all scenarios (Lewis, 1995a). Figure 1 shows the PSSUQ benchmarks established during the competitive usability benchmarking. We collected data from three products regarded as the most likely competitors of Simon. Analysis of the problems discovered during these evaluations provided additional opportunities for improved design in Simon.

Iterative Usability Studies

We conducted three fairly extensive problem discovery studies at different stages during Simon development (early 1992 prototype, first design with reasonably comprehensive function, and the design immediately preceding the final design). Our philosophy for these studies was that measurement of scenario performance and preference variables were important, but that problem discovery was more important. As long as you have competitive benchmarks, scenario measurements give you an idea about where you are relative to your competition, but provide no real guidance about what to do when your product fails to measure up. Analysis of usability problems, on the other hand, provides strong guidance for product redesign. We used the methods described in Lewis (1994b) to determine appropriate sample sizes for these studies. As a consequence of this process of iterative problem identification and design improvement, each iteration showed significant improvement in both user performance and satisfaction. Figure 1 shows the PSSUQ scale ratings for the final iteration (showing means and 95% confidence intervals), with the competitive PSSUQ benchmarks for reference. (A lower PSSUQ score is better than a higher one.) As Figure 1 shows, Simon significantly exceeded its benchmarks for all PSSUQ scales.

Icon Assessment

Most icons that appear on Simon include a descriptive label. There are four icons, however, that appear on every Simon screen. Because these icons appear on every screen, we had a design goal to provide small icons that did not require labels (conserving valuable screen space). We assessed these icons using a battery of icon assessment methods including a matching and confidence task, icon production task, and a semantic differential (Lewis, 1988; Lin, 1992). The outcome of the study indicated a problem with recognition of the icon representing access to the non-phone office tools, and led to re-representation of the function with a focus on its access to a mobile office.

Language Guidelines and Automated Readability Measures

An often neglected area of usability design and evaluation is that of language. Even modern, otherwise usable, systems often contain complicated terms for which there are much more common names. On-line messages and other documentation contain numerous sentences in the passive voice that it would be easy to recast in active voice. These considerations might seem trivial, except that psycholinguistic research has shown that (1) frequency of occurrence of a word in a language significantly affects the speed of human lexical access (Forster, 1990) and (2) it is harder to extract meaning from a passive sentence relative to its active counterpart (Bailey, 1989). To promote clarity and consistency in terminology, I provided the Simon developers with a set of language guidelines, and iteratively reviewed messages and documentation against the guidelines. Our source book for determining the best word to use when considering several synonyms was The Living Word Vocabulary (Dale and O'Rourke, 1981). I also selected random text samples from competitors' documents and developed competitive readability benchmarks for text cloudiness (a measure based on the number of specifically identified abstract words and passivized verbs in a passage divided by the number of words in the passage). At the end of Simon development, measurements taken from a random sample of texts from Simon's documentation showed that the Simon texts had a significantly lower (lower is better) text cloudiness than any of its competitors. Furthermore, using data collected during competitive usability benchmarking and iterative usability studies, Simon had a significantly better PSSUQ Information Quality rating (Lewis, 1995b) than any of its competitors.

Statistical Modeling

Because Simon had a relatively small display area, it was necessary to provide some simple statistical modeling for the size of calendar entries (Lewis, 1993a) and name lengths (Lewis, 1993b) to provide guidance to the calendar and address book developers. The calendar entry research indicated that: (1) managers use computer calendars more than non-managers; (2) managers have more entries per day than

calendar-using nonmanagers; and (3) for user-generated entries, the 95th percentile for the number of characters in an entry was 253. The name length research showed that the mean name length in the United States was about 14 characters, and that a touch-screen button that could show 20 characters would show a person's complete name 99.2% of the time (in the United States).

Designed Experiments

On occasion, it was necessary to conduct designed experiments to answer questions that arose during development. One such experiment (Lewis, 1994a) explored different screen designs for setting dates and times. Although such settings seem straightforward, users have conflicting direction stereotypes that appear to preclude the use of arrows alone for setting times and dates. Two other experiments (Lewis, Allard, and Hudson, 1994; Lewis, 1995a) evaluated different aspects of Simon's predictive keyboard. A predictive keyboard is an on-screen keyboard that contains fewer buttons than a standard keyboard, and uses linguistic probabilities to predict which letters a user will most likely want to type next. These most-likely letters appear in the keyboard's buttons. Lewis, Allard, and Hudson (1994) studied the effects of different word populations, number of displayed letters, and number of trigraph tables on the likelihood that the desired next letter would appear on the predictive keyboard. Lewis (1995a) studied input rates and user preference for the three Simon data input methods (tapping on a small on-screen standard keyboard, tapping on the predictive keyboard, and handwriting on the sketch pad). The results showed that the most effective and preferred input method was tapping on the standard keyboard. In conducting these experiments, the experimental designs described in Lewis (1993c) were quite useful.

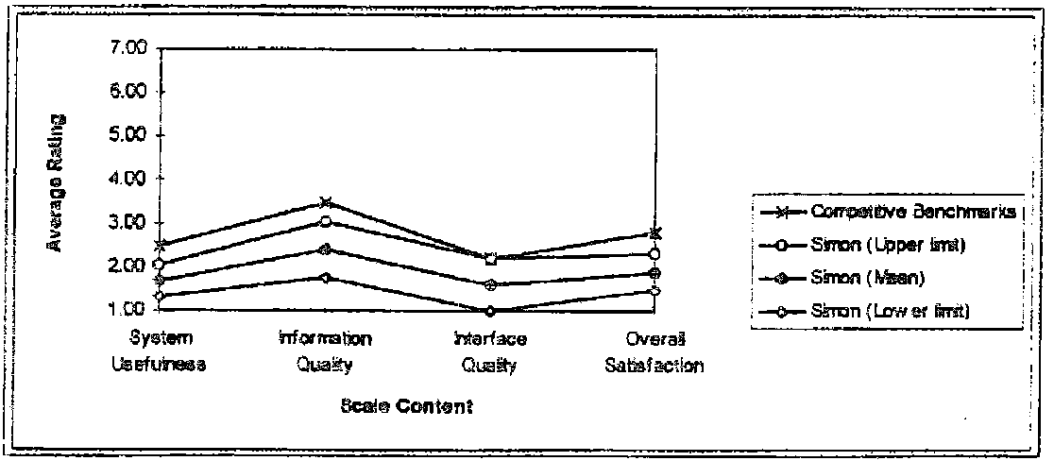

Figure 1. PSSUQ scale scores for Simon and competition

INDUSTRY RECOGNITION

One indication of the success of Simon's design is that it won the Best of Show award at Comdex '93, won an Award of Distinction in the 1994 BYTE awards (BYTE, January 1995), and was a Grand Award winner in the 7th Annual Best of What's New awards (Popular Science, December, 1994). The following quotations from reviews of Simon in trade journals also reflect the success of the usability effort.

"It looks and feels like a product you already know how to use, rather than a new religion you must immerse yourself in." (O'Malley, 1994)

"I hope that Simon is the first in a long series of personal communications tools, but even as a first generation product, Simon is a joy to use." (Nelson, 1995)

"Simon is not the first personal communicator product I've demoed, but it is by far the most comprehensive, well-designed, and easiest to use." (Carter-Lome, 1994)

DISCUSSION

This paper has described the broad range of usability evaluation methods applied to the development of Simon. The industry recognition for Simon stands as evidence for the success of the application of modern usability evaluation methods in this case. The breadth of methods also suggests that professional usability practitioners need to be fluent with a wide array of usability techniques because different development situations demand the application of different usability methods. Some of these methods come from traditional experimental psychology (statistical modeling, designed experiments, literature reviews), and others are more recent techniques (heuristic evaluations, competitive usability benchmarking, scenario-based usability problem discovery studies). All of these techniques have potential application in product development, and deserve a place in the toolbox of the professional usability practitioner.

ACKNOWLEDGMENTS

Successful product development is a team effort, and a human factors engineer is one member of a team. I gratefully acknowledge the contributions of Simon management, software development and hardware development to the final usability of Simon. Also, the efforts of Suvit Nopachai, who served with us as a graduate intern in human factors during Simon development, were invaluable.

REFERENCES

BAILEY, R. W. (1989). Human performance engineering: Using human factors/ergonomics to achieve computer system usability. (Prentice Hall, Englewood Cliffs, NJ).

BYTE. (January 1995). 1994 BYTE awards. BYTE, vol. 20, 49-60.

CARTER-LOME, M. (1994). A Simon for our times. CM: Cellular Marketing, 6.

DALE, E., and O'ROURKE, J. (1981). The living word vocabulary. (World Book, Chicago).

FORSTER, K. (1990). Lexical processing. In Osherson, D. N., and Lasnik, H. (Eds.), Language (pp. 95-131). (MIT Press, Cambridge, MA).

LEWIS, J. R. (1988). A review of symbol test methodologies (Tech. Report 54.475). (IBM Corp., Boca Raton, FL).

LEWIS, J. R. (1992). Literature review of touch-screen research from 1980 to 1992 (Tech. Report 54.694). (IBM Corp., Boca Raton, FL).

LEWIS, J. R. (1993a). Calendar entry statistics for computer calendar users (Tech. Report 54.754). (IBM Corp., Boca Raton, FL)

LEWIS, J. R. (1993b). Name length statistics for touch-screen buttons (Tech. Report 54.810). (IBM Corp., Boca Raton, FL).

LEWIS, J. R. (1993c). Pairs of Latin squares that produce digram-balanced Greco-Latin designs: A BASIC program. Behavior Research Methods, Instruments, and Computers, vol. 25, 414-415.

LEWIS, J. R. (1994a). Direction stereotypes for setting dates and times (Tech. Report 54.867). (IBM Corp., Boca Raton, FL).

LEWIS, J. R. (1994b). Sample sizes for usability studies: Additional considerations. Human Factors, vol. 36, 368-378.

LEWIS, J. R. (1995a). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. International Journal of Human-Computer Interaction, vol. 7, 57-78.

LEWIS, J. R. (1995b). Input rates and user preference for three small-screen input methods: Standard keyboard, predictive keyboard and handwriting (Tech. Report 54.889). (IBM Corp., Boca Raton, FL).

LEWIS, J. R., ALLARD, D. J., and HUDSON, H. D. (1994). Predictive keyboard design study: Effects of different word populations, number of displayed letters, and number of trigraph tables (Tech. Report 54.846). (IBM Corp., Boca Raton, FL).

LEWIS, J. R., HENRY, S. C., and MACK, R. L. (1990). Integrated office software benchmarks: A case study. In Human-Computer Interaction – INTERACT '90 (pp. 337-343). (Elsevier, London, England).

LIN, R. (1992). An application of the semantic differential to icon design. In Proceedings of the Human Factors Society 36th Annual Meeting (pp. 336-340). (Human Factors Society, Santa Monica, CA).

NELSON, M. W. (March/April 1995). The Simon personal communicator. PDA Developer, vol. 3.2, 13-16.

NIELSEN, J. (1992). Finding usability problems through heuristic evaluation. In Proceedings of the Association for Computing Machinery CHI '92 Conference (pp. 373-380). (ACM, Monterey, CA).

O'MALLEY, C. (December, 1994). Simonizing the PDA. BYTE, 145-147.

POPULAR SCIENCE. (December 1994). Best of what's new: The year's 100 greatest achievements in science & technology. Popular Science, 50-76.

SEARS, A., and SCHNEIDERMAN, B. (1989). High precision touchscreens: Design strategies and comparisons with a mouse (Tech. Report CS-TR-2268). (University of Maryland, College Park, MD).

TSOI, K. C. (April 1993). User interface issues for cellular phones. Cellular Business, vol. 10, 32-43.