Usability Problem Identification Using Both Low- and High-Fidelity Prototypes

Robert A. Virzi, Jeffrey L. Sokolov, Demetrios Karis

GTE Laboratories Incorporated
40 Sylvan Road
Waltham, MA 02254 USA
{rvirzi, jsokolov, dkaris}@gte.com

ABSTRACT

In two experiments, each using a different product (either a CD-ROM based electronic book or an interactive voice response system), we compared the usability problems uncovered using low- and high-fidelity prototypes. One group of subjects performed a series of tasks using a paper-based low-fidelity prototype, while another performed the same tasks using either a high-fidelity prototype or the actual product. In both experiments, substantially the same sets of usability problems were found in the low- and high-fidelity conditions. Moreover, there was a significant correlation between the proportion of subjects detecting particular problems in the low- and high-fidelity groups. In other words, individual problems were detected by a similar proportion of subjects in both the low- and high-fidelity conditions. We conclude that the use of low-fidelity prototypes can be effective throughout the product development cycle, not just during the initial stages of design.

Keywords

Method, usability testing, low-fidelity prototyping

INTRODUCTION

The use of low-fidelity prototyping techniques has blossomed over the last five years, with researchers claiming low-fidelity prototypes: (1) are an efficient way to search the design space [7]; (2) are predictive of preferences in the actual product [9]; (3) enhance user participation in the design process [4]; (4) enable visualization of possible design solutions [3]; and (5) provoke innovation [10]. While there have been dissenting voices to the acceptance of prototyping in general [6], current practice seems to demand some level of prototyping activity during the design process. As a profession, it is incumbent upon us to define ways to use this tool effectively, while at the same time understanding its limitations.

CHI 96 Vancouver, BC Canada
© 1996 ACM 0-89791-777-4/96/04..$3.50

In this paper, we address a critical question that appears to remain open: *In the later stages of user-interface design, are low-fidelity prototypes as effective as high-fidelity prototypes in identifying usability problems?* We cast the question this way because we think it is apparent that low-fidelity prototypes are useful in the early stages of design, before the details of the user interface are well known. One might also argue that even in later design phases, when it would be possible to build a high-fidelity working simulation of the product, low-fidelity techniques could be as effective in identifying usability problems. If this were true, then the choice of using low- or high-fidelity techniques would be based on considerations other than efficacy (e.g., cost). Alternatively, one could argue that low-fidelity prototypes can never be as effective as high-fidelity prototypes late in the design process. That is, one should build as high fidelity a prototype as possible, striving for realism as design activities progress. There are two forms this argument might take. First, one might suppose that there are particular aspects of a design that cannot be adequately simulated in a low-fidelity prototype. Thus, in using a low-fidelity prototype a usability specialist would risk missing an entire class of usability problems. A second argument against low-fidelity prototypes is that, due to their very nature, they are less effective in detecting problems; that is, they are a blunt instrument when compared to a high-fidelity prototype. Thus, use of a low-fidelity prototype would mean risking sensitivity in detecting problems. That is, any given subject would detect a smaller percentage of the total problems using a low-fidelity prototype compared to a high-fidelity prototype. An investigator could compensate for this, however, merely by increasing the number of subjects when using a low-fidelity prototype.

Two recent studies are relevant to the current question. Nielsen [5] compared the effectiveness of two types of prototypes in a heuristic evaluation, a low-fidelity paper mockup and a higher fidelity (but not high fidelity) computer mockup. Nielsen claimed that evaluators using the computer mockup were more likely to find the major problems than evaluators using the paper mockup, but the study is difficult to evaluate because he provides very few procedural details. Some of the differences between mockups could have arisen from evaluator variability, and the most important result, the difference between conditions

in finding major problems, was based on a measure of problem severity that was not validated.

Wiklund et al. [9] examined the relationship between the aesthetic refinement of a prototype and perceived usability. They created four versions of an electronic dictionary that varied in how realistically they represented what the actual product looked like. Subjects rated the prototypes on a variety of scales including ease of use and ease of learning, both before and after using the prototypes. These ratings were not affected by level of realism. They also had these subjects use the actual device, and provide the same ratings. This manipulation pointed out one of the pitfalls of low-fidelity prototyping. Because they did not accurately represent the slow response times for some aspects of the actual device's performance, estimates of usability for all the prototypes were greater than that of the actual device. Wiklund et al. argue that prototype fidelity does not affect how sensitive the test is, that is, low-fidelity tools are not blunt, but it does affect the kinds of problems one can detect.

Prototype fidelity is a continuum, not a dichotomy, and a prototype can vary from the final product along several orthogonal dimensions, including breadth of features, degree of functionality, similarity of interaction, and aesthetic refinement. Breadth of features refers to the number of features the prototype supports. Each of these features can then vary in its degree of functionality, or the extent to which the details of its operation are complete. Similarity of interaction refers to how one communicates with the product (whether by pressing buttons, clicking a mouse, touching a screen, speaking, etc.), and aesthetic refinement refers to aspects of the product that do not directly influence its functionality, such as choice of colors and graphic design. Although prototype fidelity is difficult to define precisely, a prototype that compromises on one or more of these four dimensions in a way that is obvious to the user is a low-fidelity prototype.

Low fidelity prototypes may have fairly complete breadth of features and degree of functionality and so may be similar to the final product on these dimensions, but users do not typically interact with low-fidelity prototypes in the same manner as the final product, and they do not typically look and feel the same with respect to the last dimension, aesthetic refinement.

Although illuminating, the studies reviewed above have not addressed our particular issue. We wanted to know if low-fidelity prototypes are as effective as high fidelity prototypes in detecting problems under the following conditions: (1) at the later stages of the product design process (i.e., when enough is known about the application to build a high-fidelity prototype); (2) when the study is under the supervision of usability specialists; (3) when the think-aloud protocol is used [1]; and (4) when the primary measure of effectiveness is the number of usability problems uncovered in a user interface.

We present the results of two separate experiments designed to examine this question. In both experiments, we ran two usability studies, one using a low-fidelity paper prototype, and one using a high-fidelity prototype. In experiment 1, we used an encyclopedia in electronic book form. Experiment 2 is essentially a replication of experiment 1 using a different product, an Interactive Voice Response (IVR) system for a new telecommunications service. We chose these two applications because of the differences in the products and in the low-fidelity prototyping techniques we could use. The low-fidelity version of the electronic book was created using paper and index cards, while the IVR system was simulated by a person reading the prompts aloud from written specifications.

EXPERIMENT 1 METHOD

Description of Application and Prototypes

We used a portable electronic-book player running an abridged encyclopedia. The player and the encyclopedia were tested as a complete system. The device itself is capable of displaying text and limited graphics on a flip-up, backlit LCD screen approximately 2.5" x 2.5". The overall device is 6" x 4" x 2". The device includes a limited QWERTY keyboard and some dedicated function keys. The software running on the device, a highly abridged encyclopedia, allows users to search the database and retrieve articles using any of 5 different search functions. Most articles are purely text; however, a small subset of articles have associated pictures. All graphics and text appear as black lines on a white screen. The actual electronic-book served as our high-fidelity prototype.

For the low-fidelity version of the electronic book, we created a simulation of the screens and keyboard on paper. The text and pictures used for the prototype were based on the screens from the actual product. They were prepared using a computer drawing program (Claris MacDraw Pro™), and printed on index cards. As users interacted with the paper prototype, by pressing the buttons on the paper rendition of the keyboard, they called out the keys they were pressing. The experimenter then simulated the action of the actual device by changing the display on the prototype (i.e., by removing the current card and substituting the card for the next screen). Even actions as simple as moving a highlighter from one item to the next were simulated by changing index cards. Over 100 cards were prepared that could represent not only the correct sequence of actions to complete the tasks, but also the most frequent mistakes, based on pilot testing.

Subjects

Twenty college-age subjects were recruited from local universities and paid a small fee for their participation in the study.

Procedures

Three tasks were devised that exercised the functionality of

the electronic book (find the birth date of Increase Mather, a Puritan clergyman; find a list of different types of governments; find a picture of a map of Afghanistan then determine the literacy rate for that country in the early 1980s). Half the subjects (10) performed these tasks using the actual electronic book while the other half performed the tasks using the paper prototype. All subjects were instructed to think-aloud while interacting with the system. The entire session was videotaped, and these videotapes were later used for analysis. Sessions lasted from 30 to 45 minutes.

After all testing was completed the tapes were coded. We filled out a separate incident sheet for each usability problem encountered by a subject. We declared a usability problem had occurred when: (1) the user verbally indicated that something was unclear or confusing, even if they did not make an error; (2) the user's utterances indicated a misconception regarding what was happening or what function a particular button may have had; or (3) the user's actions indicated an incorrect path or course of action, even if the user was not conscious of the problem. The number of times that a problem was identified by a given subject was also indicated on the sheets.

When all the problem sheets had been prepared, we created a single list of problems by sorting the sheets into groups of behaviors that underlay the same usability problem (problem sheets from both conditions were combined). For example, even though subjects exhibited different sorts of behaviors that indicated they did not know how to use the search facility, we grouped all tokens of this problem into a single problem type. We did this for both the high- and low-fidelity subjects at the same time. It is important to note that we were blind to the condition under which each problem sheet was generated during the sorting process. After completion of the sorting task, we generated two matrices of subjects by problem type, one for each condition of the study (low- versus high-fidelity).

EXPERIMENT 1 RESULTS

A total of 38 distinct usability problems were identified over both the low- and high-fidelity conditions (e.g., item selection was difficult; there were no instructions on how to use the search facility; function keys were not always available; their placement was inconsistent; and their labels were not descriptive). The proportion of the 10 subjects in the two conditions who identified each problem was calculated. Thus, for each problem we had a measure of how likely it was to be detected using the two prototypes. These measures form the basis of our analyses.

The first question we tried to answer was, do the techniques differ in their overall sensitivity in detecting problems? If one of the techniques was more sensitive, we would expect that the proportion of subjects finding problems using that technique would be systematically higher. In the high-fidelity group, each subject found, on average, 38% of the problems, while in the low fidelity group each subject found 34% of the problems. We tested the difference between the groups using a two-tailed paired t-test, with problems forming the unit of analysis. The two conditions did not differ significantly in their overall sensitivity ($t(37) = 1.08$, $p = .29$). The most prevalent problem, which involved a key labeled "YES" instead of an "ENTER" key, was found by 80% of the subjects (70% in the high-fidelity group, 90% in the low-fidelity group). The three least prevalent problems were found by only one subject each. Only one person in the high-fidelity condition complained about not having a "backspace" key to fix typos by deleting single characters. A single subject in the low-fidelity condition had difficulty determining from screen feedback whether the page-down command had worked as intended. The last idiosyncratic problem was also found by a low-fidelity subject who had trouble determining the boundaries of the screen in the prototype.

Figure 1 plots the average proportion of the total set of usability problems one would expect to uncover as a function of the number of subjects run in the evaluation. This is calculated using the formula described in [8],

$$1 - (1 - p)^n$$

where p is the mean probability of problem identification and n is the number of subjects in the evaluation. Separate plots are presented for the two groups. The closeness of the two curves reflects the similarity in detection ability using the two prototypes.

We realize that it is impossible to prove the null hypothesis that the two techniques are equally sensitive. We present an additional analysis here in which we ask whether or not the two kinds of prototypes were finding the same sorts of problems. It is possible for the two techniques to have approximately the same mean sensitivity, while tapping into different pools of problems. If this were true, then we would expect to find a poor correlation between the conditions. On the other hand, if the two techniques tend to uncover the same problems at comparable levels of detectability, we would expect a high correlation. Figure 2 presents a scatter plot of the data, where the proportion of low-fidelity subjects finding a problem appears on the X-axis, and the proportion of high-fidelity subjects finding the same problem appears on the Y-axis. Dots represent single problems and co-occurring points are mapped to spiked icons, where the number of points in a symbol corresponds to the number of overlapping points (e.g., a bar is two coincident points, etc.).

The ability to detect problems using the two techniques is highly correlated, $r = .58$, $p < .01$. This indicates that the two prototypes were indeed tapping into the same problems with roughly the same degree of sensitivity to most problems. Problems that were found by a high proportion of subjects in the low-fidelity condition were also found by a high proportion of subjects in the high-fidelity condition, and vice versa. Of the 38 total problems, 34 were found by subjects in the high-fidelity group, and 32 were found by subjects in the low fidelity-group.

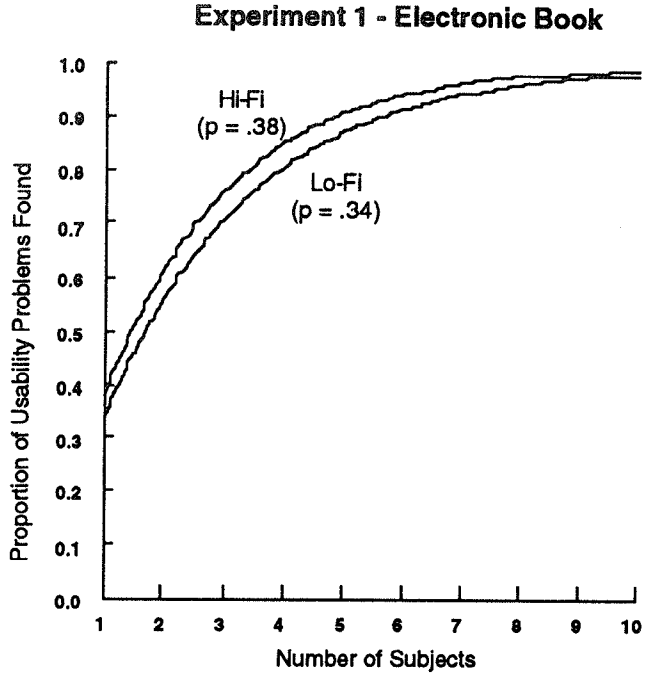

Figure 1. The proportion of usability problems one would expect to detect for a given number of subjects in the evaluation is shown for both low- and high-fidelity conditions in Experiment 1.

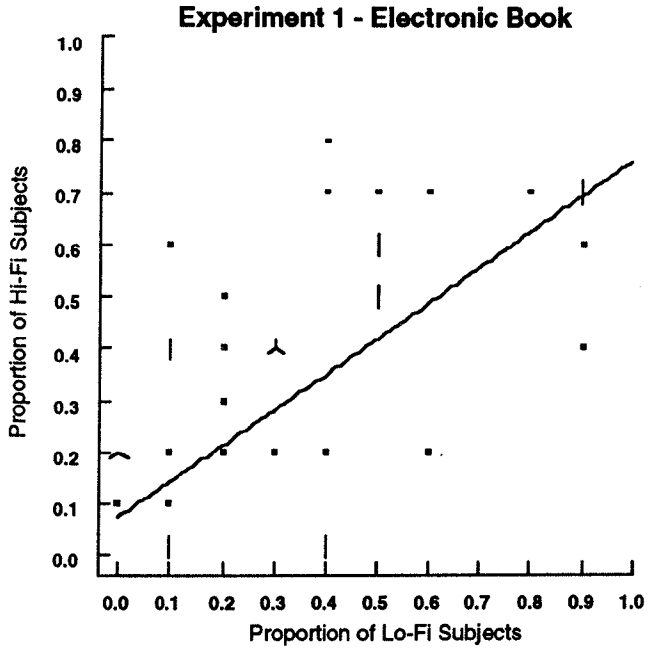

Figure 2. The proportion of subjects identifying individual problem types in the two conditions of Experiment 1 are plotted. The regression line shown corresponds to the equation: y = .075 + .690(x), r = .576.

EXPERIMENT 2 METHOD

Description of Application and Prototypes

In experiment 2, we used an Interactive Voice Response (IVR) system for a new telecommunication service. The high-fidelity prototype was built using TFLX™, a Magnum Software product for the Apple Macintosh™ computer. Users interacted with the system by calling the computer from a touch-tone phone. In response to their keypresses, the computer played back recorded digital speech via the telephone network. The application allowed callers to setup and maintain a fairly complicated telephony service that included caller identification, call screening, and automatic callback functions.

For the low-fidelity version of the IVR system, we enacted the system with one of the authors playing the part of the computer. Subjects sat in a room with the experimenter and indicated what they would do by pressing buttons on a telephone (which was not attached to the network). The experimenter read out loud what the computer would say based on the subject's input.

Subjects

Twenty college-age subjects were recruited from local universities and paid a small fee for their participation in the study.

Procedures

A series of tasks were devised that exercised the functionality of the IVR system (e.g., create a caller acceptance list; define call routing options). Ten subjects performed these tasks using the high-fidelity prototype while the other half performed the tasks using the low-fidelity prototype. All subjects were instructed to think-aloud while interacting with the system. The entire session was videotaped, and these videotapes were used for later analysis. In this experiment, the low fidelity prototype matched the high-fidelity prototype in terms of completeness; any action possible in the final system could be taken by subjects using either the low- or high-fidelity prototypes.

After user testing was completed, the tapes were coded and problem sheets were sorted as in Experiment 1, producing a subject by problem matrix for each condition.

EXPERIMENT 2 RESULTS

A total of 21 distinct usability problems were identified over both the low- and high-fidelity conditions (e.g., difficulty finding the setting for blocked calls; confusing terminology; misleading prompt order on a menu; and inconsistent key mappings). The proportion of the 10 subjects in the two conditions who identified each problem was calculated, as in experiment 1.

A t-test on these data replicated the finding from experiment 1: The two techniques did not differ significantly in their overall ability to detect usability problems ($t(20) = .870$, $p = .39$). In the high-fidelity group, each subject found, on average, 40% of the problems, while in the low-fidelity group subjects found 46% of the problems. Figure 3 shows these results graphically.

Figure 4 presents a scatter plot of the data for experiment 2. As in experiment 1, the ability to detect problems using the two techniques is correlated, $r = .49$, $p < .05$. This indicates that the two prototypes were indeed tapping into the same problems with roughly the same degree of sensitivity to most problems. Of the 21 total problems, 19 were found by subjects in the high-fidelity group, and 20 were found by subjects in the low-fidelity group. Two problems appear to be outliers in that they were identified by almost all the subjects in the high-fidelity condition but relatively few of the subjects in the low-fidelity condition (see Figure 4). The two problems were related in that they were both caused by the failure of the high-fidelity prototype to automatically save users' changes upon returning to the main menu from a sub-menu. The low-fidelity prototype handled this function slightly differently, thus subjects were not as likely to detect the problem.

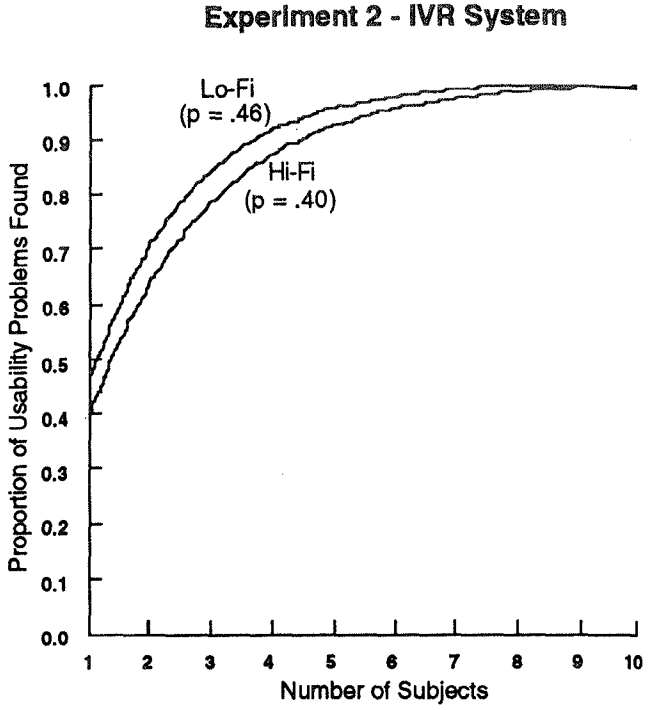

Figure 3. The proportion of usability problems one would expect to detect for a given number of subjects in the evaluation is shown for both low- and high-fidelity conditions in Experiment 2.

DISCUSSION

Are Low-Fidelity Prototypes a Blunt Tool?

The first question we attempted to answer was, *Are low-fidelity prototypes a blunt tool?* The results of both experiments were highly consistent, and indicated that low-fidelity prototypes are as effective as high-fidelity prototypes at detecting usability problems. Thus low-fidelity prototyping is not a blunt tool, or at least it is as sharp as high-fidelity prototyping in detecting usability problems. We employed a usability technique involving a think-aloud protocol, and we are not claiming that our results would apply equally well to other usability techniques. The types of low-fidelity prototypes used here require extensive manipulation by the experimenter as well as frequent interaction between experimenter and subject. The think-aloud technique also requires interaction, as the experimenter sometimes needs to ask questions to understand the problems a subject is experiencing, and this technique is a nice complement to the use of low-fidelity prototypes.

While we acknowledge that one cannot prove the null hypothesis (that there is no difference between prototyping techniques), we would hasten to point out the practical importance of the finding. Both studies showed that substantially the same sets of problems were found in the low- and high-fidelity groups. Thus a user-interface designer would have essentially the same amount of information regardless of the technique employed. Note that low fidelity does not imply a lack of functionality. In the IVR system the functionality of the prototypes in both conditions were almost identical, and complete. For the low-fidelity version of the electronic book, just enough

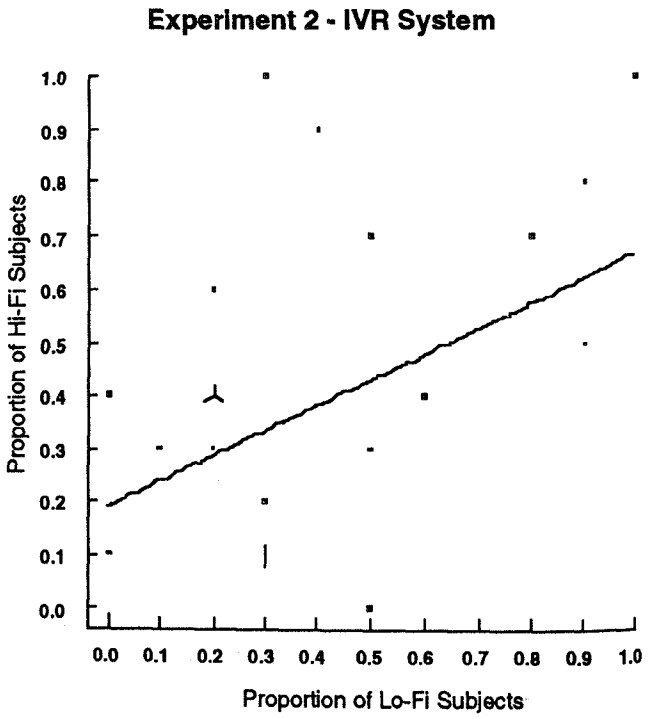

Figure 4. The proportion of subjects identifying individual problem types in the two conditions of Experiment 2 are plotted. The regression line shown corresponds to the equation: $y = .181 + .480(x)$, $r = .487$.

functionality was provided to allow the user to perform the tasks. The electronic book itself was obviously fully functional.

Although the products in the two experiments were quite different, and used different low-fidelity prototyping techniques, the results were essentially the same. How generalizable are these results? Surely they would not apply, we have heard colleagues say, to direct manipulation interfaces, virtual reality or other immersive systems, or to systems that are extremely response-time sensitive [2]. We are not so sure. The exact bounds of our results must await additional research with a variety of products. In a third experiment (not reported here), we tested a complex, CD-ROM, multimedia Macintosh™ product that had a graphical user interface. Our preliminary analysis finds very few differences in the usability problems discovered using a paper-based low-fidelity prototype versus the actual product.

Do Low-Fidelity Prototypes Find Different Problems?

Our second question was, *Do low-fidelity prototypes tend to find different problems than high-fidelity prototypes?* The question appears to be answered with a qualified, *No*. Both techniques uncovered the same usability problems for the most part, and at about the same level of sensitivity, as witnessed by the high positive correlations in the two studies between the proportion of low- and high-fidelity subjects finding particular problems. That is, when a problem was found by a high proportion of subjects in the high-fidelity group, it also tended to be found by a high proportion in the low-fidelity group. However, some problems that were found in the high-fidelity condition were not found in the low-fidelity condition, and vice versa. This is true for both studies. There are many reasons why this result might obtain, but inspection of the actual problems did not lead to any obvious conclusions. We believe, that except as noted earlier, this is probably a result of normal variability, given the small number of subjects, rather than a function of differences between the two prototypes, but we currently have no evidence to substantiate this belief.

Would there have been differences between the two conditions if we had considered problem severity? It is very difficult to come up with adequate techniques for defining and labeling problem severity, and we did not attempt to do so in these experiments. However, based on an "eye-ball" examination of the problems uncovered, we could see no striking differences as a function of our rough estimates of problem severity. Both groups found many severe problems, as well as many we would categorize as minor. Where there were large differences between the two conditions, it was difficult to attribute this to aspects of the prototypes. For example, in experiment 1, problem 32 involved using the graphic search capability: an instruction screen was presented showing three screen locations where a cursor could be located to navigate through graphics screens or select the associated article. Not only was this difficult to understand as presented, but this information had to be remembered, because when the graphics appeared all instructions disappeared. This resulted in a variety of related problems that we categorized together. In the low-fidelity group, 90% of the subjects experienced these problems, while only 40% experienced them in the high-fidelity group. We just don't know whether there is some reason for this difference related to how the subjects performed the task in the two groups, or whether this is random variability. It was clear after watching only one or two subjects struggle with this problem that it was quite serious. Problems that were found by only one or two subjects did not appear to be as serious, but we have no real data to support this claim.

It is also important to remember that practically all problems were identified within each condition (i.e., across subjects). In experiment 1, the high- and low- fidelity groups uncovered 34 and 32 of the 38 problems, respectively, while in experiment 2, the numbers were 19 and 20 out of 21.

In both experiments we set up and performed the usability evaluation as realistically as possible, and chose what we considered a representative set of tasks that exercised all the main functions of the product. In fact, the IVR system in experiment 2 was part of a new telecommunication service that GTE planned to offer, and the high-fidelity prototype was built to support our user-interface design effort. It was built from specifications we prepared, and used for usability testing essentially identical to that reported here. That is, the same tasks, instructions, and procedures used in usability tests for the commercial product were used in the usability tests reported here. The high-fidelity prototype provided full functionality and was quite complex; it contained over 5,000 TFLX™ "icons," seven databases, over 500 sound files, and took over two months to develop. As we note below, this prototype served several purposes. However, with respect to usability testing, had we performed an experiment like the one reported here a year earlier, *we would not have spent the time and effort to build a high-fidelity prototype, and would have conducted all usability testing with the low-fidelity prototyping techniques reported here.* In fact, this is how we currently design IVR systems in practice.

In experiment 1, the commercial electronic-book product served as the high-fidelity prototype. This is actually a conservative approach because a high-fidelity prototype, by definition, can equal, but never surpass, the fidelity of the actual product. Using the actual product instead of a high-fidelity prototype would tend to increase, not decrease, the differences between our groups if the high-fidelity prototype failed to incorporate some of the features or attributes of the real product.

The Need for High-Fidelity Prototypes

Although we found substantially the same set of usability problems in both the low and high-fidelity groups, we are not arguing that there is no place in the development cycle for high-fidelity prototypes. There are several

circumstances in which high-fidelity prototypes are useful. First, there will probably be a set of usability problems for which any particular low fidelity prototype will be inadequate. For example, in experiment 1 there were no tasks requiring the user to physically manipulate the electronic book, which would be required when opening it to insert a diskette. It would not have been possible to complete this task given the two dimensional paper and cardboard low-fidelity prototype. We should note, however, that a low fidelity prototype that allowed physical manipulation could have been constructed. Low-fidelity prototypes will always be limited, by definition, on some dimensions; however, given a particular set of tasks, it is almost always possible to construct a low fidelity prototype that can support their execution. In a low-fidelity prototype of an IVR system, as in our second experiment, it is not possible to check for possible problems with the concatenation of prompts, or their intelligibility (especially, for example, if synthetic speech were to be used) because the experimenter reads the prompts from the specification. However, even with a high-fidelity prototype of an IVR system, it is not typically possible to check concatenation or intelligibility unless special care was taken to build the prototype to match the commercial product (e.g., by breaking up and concatenating messages in the same way, or using the same text-to-speech synthesizer).

Even though we are focusing here on the prototype itself, it may be helpful to step back and consider what is happening from a broader perspective. The task in a usability evaluation is to identify problems. When this is done in the laboratory, even when using a high-fidelity prototype, we are simulating not only the product, but also the tasks, the environment, the data used by the product, and so on. Thus even a high-fidelity prototype can lead to a noisy approximation of the actual way in which people will use a product. The focus of this paper is the artifact (the prototype), but it is also important to focus attention on these other aspects of testing as well.

Another general class of problems that may be identified differentially in low- and high-fidelity prototypes involve performance measures such as the time to complete certain operations. In many cases, of course, it takes longer for an experimenter to manipulate a low-fidelity prototype in response to a user's action than for a computer running a high-fidelity prototype to respond. We are currently quantifying the magnitude of these differences in another experiment that measures time to complete tasks and other performance measures, including the number of tasks completed and the number of steps required per task.

Once we leave the realm of usability, there are a variety of areas in which increasing the fidelity of certain aspects of the prototype will prove much more useful. In selecting the fonts, images, and colors to use in a graphical user interface, for example, a prototype that can provide displays nearly identical to the actual product will be far superior to a paper simulation.

In our development work, we have found three broad areas, apart from usability testing, where high-fidelity prototypes are valuable. First, prototypes are effective in communication with marketing departments, and here the high face validity of high-fidelity prototypes can be critical. Both positive and negative information about a product can be communicated effectively, either via videotapes of usability sessions or via direct interaction by marketing personnel. Marketing can also use the prototype as part of early demonstrations to their clients. The second area involves communication with developers. A prototype, as an adjunct to a user-interface specification, can be very effective in conveying a sense of system behavior, and in helping the development team to read and understand a detailed specification. A corollary to this is that the process of building a high-fidelity prototype always helps to identify weaknesses and omissions in a user-interface specification. The third area involves communication with people writing documentation or preparing training materials. Since technical writers often need to start their work long before a working version of the product is available, having access to a prototype can be valuable in helping to develop an understanding of exactly how the product works, as well as providing easy access to details such as specific prompt wording or screen designs.

Implications for Practitioners

We interpret these results to mean that a designer or usability tester need not consider sensitivity when selecting a method for representing an interface in a test, provided comparable functionality is maintained. We would condition this recommendation to the case where the designer is seeking to identify problems and is willing to use the think-aloud protocol. Given these conditions and the limited number of subjects typically used in an evaluation, the designer is likely to uncover about as many problems at any one level of fidelity as another. Moreover, using a carefully constructed low-fidelity prototype, the designer should be uncovering the same types of problems as if a high-fidelity prototype were used.

ACKNOWLEDGMENTS

We would like to thank Devorah Klein for her assistance in conducting these studies. This work has also benefited from comments and discussions with department members, notably David Fay and Greg Cermak. Finally, we would like to thank Larry Wood, Alan Asper, Frank Reiff, Joe Dumas, Jonas Lowgren, and other members of the U-TEST mailing list for comments on a draft of this paper, as well as the CHI '96 reviewers.

MacDraw Pro is a registered trade mark of Claris, TFLX is a registered trade mark of Magnum Software, and Macintosh is a registered trademark of Apple Computer.

REFERENCES

1. Lewis, C. Using the "thinking- aloud" method in cognitive interface design. IBM Technical Report RC

9265 (#40713), 2/17/82.

2. Lowgren, J. Personal communication, September, 1995.
3. Moggridge, B. Design by story-telling. Applied Ergonomics, 24(1), (1993) pp. 15 - 18.
4. Muller, M. PICTIVE - An exploration in participatory design. In Proc. of the Conference on Human Factors in Computing Systems. ACM:New York, (1991) pp. 225 - 231.
5. Nielsen, J. Paper versus Computer Implementations as Mockup Scenarios for Heuristic Evaluation. In Proceedings of IFIP INTERACT'90: Human-Computer Interaction, (1990) pp. 315 - 320.
6. Thimbleby, H. Failure in the technical user-interface design process. Computers and Graphics, 9, (1985), pp. 187 - 193.
7. Virzi, R. What can you learn from a low-fidelity prototype? In Proc. of the Human Factors Society 33rd Annual Meeting, HFES: Santa Monica, CA (1989), pp. 224 - 228.
8. Virzi. R. Refining the test phase of usability evaluation: How many subjects is enough? Human Factors, 34(4), (1992), pp. 457 - 468.
9. Wiklund, M., Thurrott, C., and Dumas, J. Does the fidelity of software prototypes affect the perception of usability? In Proc. of the Human Factors Society 36th Annual Meeting, HFES: Santa Monica, CA, (1992), 399 - 403.
10. Wulff, W., Evensen, S., and Rheinfrank, J. Animating interfaces. In CSCW '90 Proc., ACM: New York, (1990), pp. 241 - 254.