

HIGH-FIDELITY OR LOW-FIDELITY, PAPER OR COMPUTER? CHOOSING ATTRIBUTES WHEN TESTING WEB PROTOTYPES

Miriam Walker, Leila Takayama, and James A. Landay
{mwalker | leila | landay}@cs.berkeley.edu
Group for User Interface Research, Computer Science Division
University of California, Berkeley

Interface designs are currently tested in a mixture of fidelities and media. So far, there is insufficient research to indicate what level of fidelity and media will produce the best feedback from users. This experiment compared user testing with low- and high-fidelity prototypes in both computer and paper media. Task-based user tests of sketched (low-fidelity) and HTML (high-fidelity) website prototypes were conducted in each medium, separating the testing medium from other factors of prototype fidelity. We found that low- and high-fidelity prototypes are equally good at uncovering usability issues. Usability testing results were also found to be independent of medium, despite differences in interaction style. Designers should choose whichever medium and level of fidelity suit their practical needs and design goals, as discussed in this paper.

A prototype is a working model built to develop and test design ideas. In web and software interface design, prototypes can be used to examine content, aesthetics, and interaction techniques from the perspectives of designers, clients, and users. Usability professionals often test prototypes by observing users as they perform tasks typical of the intended use of the product. By gathering data on user mistakes and comments, designers and usability professionals can find usability problems at an early stage of design, before substantial resources are invested in flawed designs.

Often web and software designers make prototypes with more than one technique, moving closer to the final production methods as the design progresses to completion (Newman & Landay, 2000). Prototypes more similar to the final product are “high-fidelity” while those less similar are “low-fidelity.” A high-fidelity prototype is often made with the same methods as the final product and hence has the same interaction techniques and appearance as the final product but is more expensive and time-consuming to produce than a low-fidelity prototype.

We investigated the affect of fidelity on user testing because if usability testing on low-fidelity prototypes is equally good to testing on high fidelity prototypes, then cheap, quick low-fidelity prototyping techniques could be employed through more of the design process. Since low-fidelity prototyping is now possible on computers as well as paper by using applications such as DENIM (Lin et al., 2000), SILK (Landay & Myers, 2001), and PatchWork (van de Kant et al., 1998), studies of the effect of prototyping fidelity should also address the dimension of medium. Hence we extended previous work by separating out the *medium* of prototype presentation from other aspects of *fidelity*, and measuring the effect on user feedback during testing. We used standard industry practices, such as think-aloud protocols and task-based user testing (Nielsen, 1994) to test the effect of fidelity and medium on prototypes.

Previous studies of usability testing combined medium and fidelity by testing high-fidelity computer prototypes against low-fidelity paper or computer prototypes, but they

did not address all combinations of paper and computer media and low and high fidelities. At the end of testing, low- and high-fidelity prototypes were found to elicit equal amounts of user feedback (Virzi, Sokolov, & Karis, 1996), but users gave different suggestions (Hong, Li et al., 2001; Virzi, 1989). We wanted to test whether prototyping medium (computer or paper) would change users’ interactions with or expectations of prototypes and hence the usability testing issues they raised.

Prototyping technique must also be balanced against practical considerations, including availability of prototyping tools, need for remote usability testing, and the effect on designers’ practices. For example, low-fidelity sites can be created on paper or on computer, but only computer prototypes can take advantage of software tools such as WebQuilt (Hong, Heer et al., 2001) that track user paths through websites.

This paper examines the relative advantages of low- and high-fidelity prototypes presented on paper or computer for user testing and reviews other considerations in the choice of fidelity and medium. We also extend previous studies by measuring differences in the number, severity, and type of usability issues; as well as proportion of the website that is the focus of participants’ comments. A good prototyping technique is one that finds the maximum number of real usability problems during user testing, while being both inexpensive and flexible for designers.

THE FIDELITY OF PROTOTYPES

Fidelity describes how easily prototypes can be distinguished from the final product and can be manipulated to emphasize aspects of the design. For example, Greeked text can be a placeholder for real content to test interpretations of different layouts in a low-fidelity prototype (Wong, 1992). Low-fidelity representations, such as sketches, differ from the final product in interaction style, visual appearance, and/or level of detail. Sketching is quick, leaving more time to iterate on designs between, or even during, usability tests. Quick, low-fidelity prototyping techniques can allow designers and users

to focus on high-level interaction design and information architecture, rather than on details or visual style (Black, 1990; Landay & Myers, 2001; Wong, 1992).

However, despite the advantages of low-fidelity prototypes, designers may move to high-fidelity prototypes early if they believe clients will judge low-fidelity prototypes as unprofessional (Newman & Landay, 2000). In addition, high-fidelity prototypes offer more realistic interactions and are better at conveying the range of design possibilities. High-fidelity prototyping, however, may make designers reluctant to change designs and less likely to fully explore the design space (Goel, 1995; Wong, 1992).

Previous studies have addressed potential differences between low- and high-fidelity without manipulating the medium of the prototypes. There was no difference in the number of usability problems found by testing with low- and high-fidelity prototypes (Virzi, Sokolov, & Karis, 1996), and aesthetics did not influence perceptions of usability (Wiklund, Thurrott, & Dumas, 1992). These experiments did not look at the potential effects of medium. Hong, Li, Lin and Landay (2001) manipulated the fidelity of computer prototypes and found no differences in user prioritization of proposed areas for design improvement. However their users rated the prototypes as different in “professionalism,” “finishedness” and “likeliness to change.”

THE MEDIUM OF PROTOTYPES

Choosing either paper or computer as the medium for a prototype has implications for the realism of the representation, the types of usability testing methods available, and the ability of users to participate in the design process. Because paper cannot respond to a mouse or keyboard, paper and computer prototypes support testing different parts of the interaction. For example, tests using low-fidelity prototypes sometimes rely on a human or computer faking the behavior of a fully working system to demonstrate the interaction. On paper a usability tester manipulates screens/sheets of paper in response to the user’s behavior (Rettig, 1994). Consequently, paper prototypes allow ad-libbed changes for exploring interactions while sacrificing some realism. Participatory design on paper is also more accessible to novices because most people can sketch (Erickson, 1995).

There are advantages and disadvantages to prototyping on computers. Some methods of computer prototyping require that the interaction flow be decided well before user testing. They allow pre-programmed responses to user behavior and make it easy to record user actions remotely. But high-fidelity computer prototypes produced using prototyping tools may also reduce design effectiveness because programming languages, HTML (Vaidyanathan, Robbins, & Redmiles, 1999), or multimedia development tools can limit designs to standard widgets and slow down the prototyping process. Computer prototyping tools may require designers to specify more implementation detail than they need or want, interrupting creative flow. Sketch-based, low-fidelity computer tools allow quick prototyping and require less skill than high-fidelity prototyping tools but these tools may further limit the range of interaction techniques available in the prototype.

METHODOLOGY

We ran an experiment in which participants carried out tasks on two different websites under two different fidelity conditions (low-fidelity and high-fidelity) and in two different media (paper and computer). We counted the number of comments made and the number and type of usability issues identified by participants.

Experimental Methodology

Our experimental methodology applied standard industry practices for usability testing within the framework of a rigorously controlled experimental design. We used think-aloud, task-based testing with users who had a range of experience and we counterbalanced our conditions for learning effects.

Participants. The 28 participants were recruited using notices posted around the UC Berkeley campus, in local stores and libraries, and by word of mouth. The experimenters knew none of the participants prior to user testing. None of the participants had professional website development experience and 61% had used online banking, the application domain we tested. Participants received \$15 in either book or ice cream vouchers in return for their time.

Apparatus. The computer medium versions of the websites were displayed on 17-inch monitors, with browser windows set to 11"x 8.5" (A4) size. The paper medium versions were printed/photocopied onto A4 paper. Participants’ comments were recorded using a digital voice recorder.

Setting. Experiments were conducted in a quiet computer laboratory in Soda Hall (the home of the Computer Science Division) on the UC Berkeley campus.

Procedure. We tested the independent variables of (1) fidelity and (2) medium using a factorial design experiment. The two-by-two design counterbalanced low- and high-fidelity, computer and paper media (see Table 1). Participants performed five typical online banking tasks: setting up an account, e-mailing themselves a checking account statement, setting up automatic bill payments, finding the value of foreign currency, and transferring money between accounts. We recorded participants thinking aloud and asked them to make additional comments at the end of each task. Designers in industry use think-aloud user testing on both low- and high-fidelity prototypes (John & Marks, 1997). Behavioral data (e.g., confusion or frustration) was noted and discussed with users at the end of each task, as were unusual or incorrect paths through the website. Each participant tested the two sites in the same fidelity but on different media (see Table 1).

Table 1. Number of participants in each condition.

	Order of Site Version & Medium		People
Low-fidelity (16 people)	Site 1 (computer)	Site 2 (paper)	4
	Site 1 (paper)	Site 2 (computer)	4
	Site 2 (computer)	Site 1 (paper)	4
	Site 2 (paper)	Site 1 (computer)	4
High-fidelity (12 people)	Site 1 (computer)	Site 2 (paper)	3
	Site 1 (paper)	Site 2 (computer)	3
	Site 2 (computer)	Site 1 (paper)	3
	Site 2 (paper)	Site 1 (computer)	3

Prototyping Methodology

For the purpose of the experiment two websites were designed with approximately the same content and functionality, but different information architectures and visual designs (see Figure 1). The experimental sites were based on online banking at small banks such as Premier Bank and The State Bank of Alcester. Nielsen's usability heuristics (1994) were used to insert a wide range of usability problems in the prototypes. The low-fidelity versions were sketched on paper whereas the high-fidelity versions were created in HTML. To make low-fidelity computer versions, the paper sketches were scanned and used as backgrounds to HTML and JavaScript pages. Hot-spots over the sketched links and forms over the sketched text-entry boxes provided realistic interaction. To make high-fidelity paper versions, the HTML web-pages were printed out. On paper prototypes, participants wrote in text boxes and tapped links, buttons and drop-down menus with a pencil. The high-fidelity, computer-presented sites had color, fonts, and images, but the information architecture was consistent between media. The high-fidelity, paper prototypes were in color. Participants used the keyboard and mouse with computer prototypes. For the complete low- and high-fidelity computer prototypes, see:

<http://guir.berkeley.edu/prototypefidelity>

Analysis Methodology

Comments made by the participants were categorized into usability issues by the researchers and were then rated by ten outside judges. The judges rated the degree to which usability problems impeded users of the website (severity) and categorized the issues into types using Nielsen's usability heuristics (Nielsen & Mack, 1994).

Transcribing and Extracting Comments. The transcriptions were reduced to a series of comments that indicated: (1) like or dislike for aspects of the website, (2) problems navigating through the site to complete the assigned tasks, (3) suggestions for site improvements, or (4) confusion about the site. Pieces of contextual information (observations recorded by the researchers) were added to some comments to indicate participants' activity at the time.

Categorizing Issues. Finding the maximum number of unique, severe usability issues is more important than the number of comments a user makes. Hence many usability studies categorize comments into usability issues (John & Marks, 1997; Virzi, Solokov, & Karis, 1996). Likewise we grouped comments into issues to compare unique usability problems across experimental conditions. For example, we can compare paper and computer media for the number of

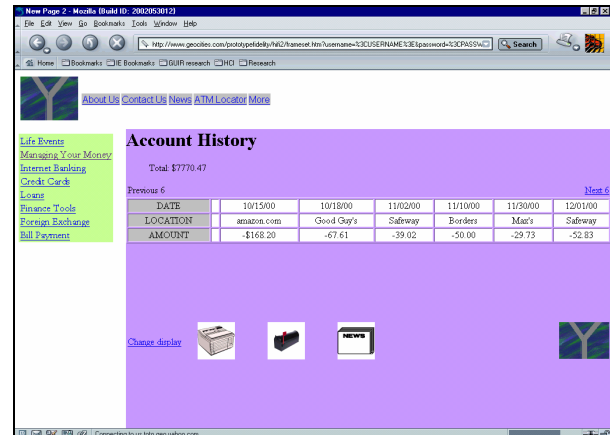
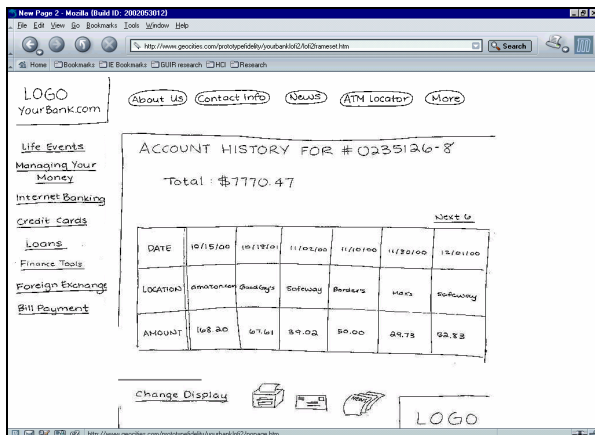
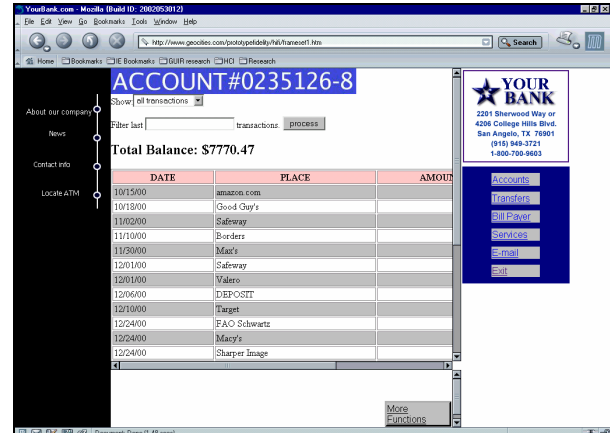
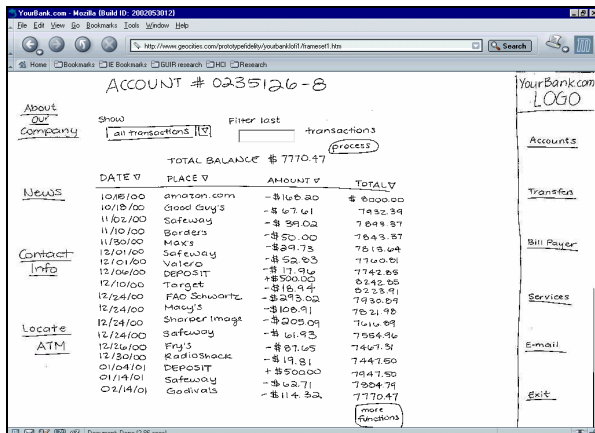


Figure 1. Account history pages for the two websites in both low-fidelity and high-fidelity. Low-fidelity websites are on the left and high-fidelity on the right. The top row is website 1 and the bottom row is website 2.

users raising issues about the lack of a help function. Finding a difference between media would tell us how best to test awareness of the absence of a help function.

Comparing Issue Types, Severity and Scope. To compare issues across experimental conditions, ten usability experts (usability academics and professionals) categorized issues according to Nielsen’s heuristic evaluation categories and by the severity of the problem. The expert judges also rated the scope of the website to which the comments applied (e.g., widget, page, whole site). Scope ratings can help to determine if high-fidelity prototypes cause usability testing participants to focus on aesthetics or other visual details.

RESULTS

By taking two measures under different conditions but from the same person we were able to minimize error in our results. Comments and issues were analyzed using the related-measures variable of medium (same person, two conditions) and the independent-measure variable of fidelity (two people, two conditions).

The 28 participants made 1270 comments ($M = 45$, $SD = 18$). On average, participants made five more comments about computer prototypes than paper prototypes (Wilcoxon Signed Ranks, $Z = -2.437$, $p = 0.015$). There was no significant difference between low- and high-fidelity conditions (Mann-Whitney, $Z = -1.463$, ns) in the number of comments.

Each participant averaged 34.6 ($SD = 11.1$) distinct usability issues of 169 identified by participants. There was no significant difference between low- and high-fidelity (Wilcoxon Signed Ranks, $Z = -1.151$, ns) or paper and computer media (Mann-Whitney, $Z = -0.186$, ns). There was no interaction effect between medium and fidelity (Chi Square, $\chi^2 = 1.233$, ns).

Participants found equally severe issues between media (Wilcoxon Signed Ranks, $Z = -0.649$, ns) and fidelities (Mann-Whitney, $Z = -1.118$, ns). Severity rating and number of comments on an issue correlated ($r = 0.401$, $p = 0.0001$).

Raters categorized issues to examine the types of issues raised and measure whether the types of issues differed between prototypes. The largest number of issues violated the usability heuristics of “match between system and the real world” (28%) and “visibility of system status” (17%). Only 10% of issues broke the “aesthetic and minimal design” heu-



Figure 2. Scope of usability issues found.

ristic, suggesting users did not focus on aesthetics. When measuring differences in types of usability problems based on comments we found no differences in the types of usability problems raised by test participants between paper and computer media (Chi Square, $\chi^2 = 4.834$, ns) or low-and high-fidelity conditions (Chi Square, $\chi^2 = 4.834$, ns). When measuring based on issues, we found differences in the types of usability problems when problems were categorized by fidelity (Chi Square, $\chi^2 = 30.70$, $p < 0.01$) but not by medium (Chi Square, $\chi^2 = 13.14$, ns).

Users did not focus on any particular level of the site (see Figure 2) and there were no differences by medium (Mann-Whitney, $Z = -0.288$, ns) or fidelity (Wilcoxon Signed Ranks, $Z = -1.555$, ns).

DISCUSSION

We found few differences between computer and paper media or low and high fidelities and therefore recommend choosing the medium and fidelity based on practical considerations of prototyping and usability testing. Quick iterations and modifications are necessary for early-stage design and are made easier in low-fidelity prototypes, on paper or computer. Because our tests were conducted under well-controlled conditions our results should generalize to less strict conditions.

Users made significantly more comments about computer than about paper prototypes but there were no differences in the number of usability issues. This suggests participants were more verbose on computer but the computers make it no easier to find usability problems. In practice, the additional comments from the computer condition may help interpret and solve usability problems. The correlation between severity and number of comments suggests severe problems are identified by more users.

The types of usability issues found were significantly different between low- and high-fidelity conditions. A limitation of the data analysis technique is that it does not allow us to attribute the difference to any particular type of issue. Anecdotally, the few comments on aesthetics were addressed to low-fidelity, not high-fidelity, prototypes. Since there were so few comments addressed to aesthetics, it is possible that task-based usability testing focuses users’ attention on task-related information and navigation, rather than visual details.

It could be argued that some results are not statistically significant because we had only 28 participants. However, such small sample sizes are actually larger than industry recommendations for usability testing of prototypes (Nielsen & Mack, 1994).

CONCLUSIONS

This user testing experiment found few differences between computer and paper or low- and high-fidelity prototypes, in the number, type, and severity of usability issues found. Our results are consistent with previous studies comparing low- and high-fidelity prototypes. Our findings of no difference support the idea of using low-fidelity prototypes for design and testing. Low-fidelity prototypes have big advan-

tages in cost and ease of iteration, and allow designers to focus on interaction design and information architecture. We found paper and computer media to be equally valid for testing prototypes. Prototyping on paper eases participatory design and enables testing in a more exploratory, dynamic way. Computer prototypes allow automatic recording of user tests, can be distributed electronically, and can help document the design process. Designers can choose the most practical prototyping medium because user-testing feedback is equally good with either.

ACKNOWLEDGEMENTS

Thanks to the members of the Group for User Interface Research for their help and encouragement, particularly Corey Chandler.

REFERENCES

- Black, A. (1990). Visible planning on paper and on screen: The impact of working medium on decision-making by novice graphic designers. *Behaviour & Information Technology* 9 (4): 283-296.
- Erickson, T. (1995). Notes on Design Practice: Stories and Prototypes as Catalysts for Communication. *Scenario-Based Design: Envisioning Work and Technology in System Development*. New York: Wiley & Sons.
- Goel, V. (1995). *Sketches of Thought*. Cambridge: The MIT Press.
- Hong, J. I., Li, F. C., Lin, J., & Landay, J. A. (2001). End-User Perceptions of Formal and Informal Representations of Web Sites, *Adjunct Proceedings of CHI 2001*, Seattle, WA, 385-386.
- Hong, J. I., Heer, J., Waterson, S., and Landay, J. A. (2001). WebQuilt: A Proxy-based Approach to Remote Web Usability Testing. *ACM Transactions on Information Systems* 19(3): 263-285.
- John, B. E. & Marks, S. J. (1997). Tracking the Effectiveness of Usability Evaluation Methods. *Behaviour and Information Technology*, 16 (4/5), 199-203.
- Landay, J. A. & Myers, B. A. (2001). Sketching Interfaces: Toward More Human Interface Design. *IEEE Computer* 34 (3): 56-64.
- Lin, J., Newman, M. W., Hong, J. I., & Landay, J. A. (2000). DENIM: Finding a Tighter Fit Between Tools and Practice for Web Site Design. *CHI Letters* 2 (1): 510-517.
- Newman, M. W. and Landay, J. A. (2000). Sitemaps, Storyboards, and Specifications: A Sketch of Web Site Design Practice. *Proceedings of Designing Interactive Systems*, New York, 263-274.
- Nielsen, J. (1994) *Usability Engineering*. San Francisco: Morgan Kaufmann.
- Nielsen, J. and Mack, R.L. (Eds.) (1994). *Usability Inspection Methods*. New York NY: John Wiley and Sons.
- Rettig, M. (1994). Prototyping for Tiny Fingers, *Communications of the ACM* 37 (4): 21-27.
- Vaidyanathan, J., Robbins, J. E., & Redmiles, D. F. (1999). Using HTML to Create Early Prototypes. *Extended Abstracts of CHI '99*, Pittsburg, PA, 232-233.
- van de Kant, M., Wilson, S., Bekker, M, Johnson, H., & Johnson, P. (1998). PatchWork: A Software Tool for Early Design. *CHI '98 Summary*, Los Angeles, CA, 221-222.
- Virzi, R. (1989). What Can You Learn From a Low-fidelity Prototype? *Proceedings of the Human Factors Society 33rd Annual Meeting*, Santa Monica, CA, 224 -228.
- Virzi, R. A., Sokolov, J. L., & Karis, D. (1996). Usability problem identification using both Low- and High-Fidelity Prototypes. *Proceedings of ACM CHI '96*, Vancouver, British Columbia, Canada, 236-243.
- Wiklund, M., Thurrott, C., & Dumas, J. (1992). Does the Fidelity of Software Prototypes Affect the Perception of Usability? *Proceedings of the Human Factors Society 36th Annual Meeting*, HFES, Santa Monica, CA, 399-403.
- Wong, Y.Y. (1992, May). Rough and Ready Prototypes: Lessons from Graphic Design. *Short Papers Proceedings of CHI '92*, Monterey, CA, 83-84.