A Comparison of Pseudo-Paper and Paper Prototyping Methods for Mobile Evaluations

Lumsden, Joanna; MacLean, R.

National Research Council Canada
Institute for Information Technology

Conseil national de recherches Canada
Institut de technologie de l'information

NRC · CNRC

*A Comparison of Pseudo-Paper and Paper Prototyping Methods for Mobile Evaluations **

Lumsden, J., and MacLean, R.
November 2008

* Proceedings of the International Workshop on MObile and NEtworking Technologies for social applications (MONET'2008), part of the LNCS OnTheMove (OTM) Federated Conferences and Workshops. November 4, 2008. Monterrey, Mexico, pp. 538-457.

A Comparison of Pseudo-Paper and Paper Prototyping Methods for Mobile Evaluations

Joanna Lumsden and Ryan MacLean

National Research Council of Canada, IIT e-Business, 46 Dineen Drive, Fredericton, N.B., Canada E3B 9W4
jo.lumsden@nrc-cnrc.gc.ca

Abstract. The research presented in this paper is part of an ongoing investigation into how best to support *meaningful* lab-based usability evaluations of mobile technologies. In particular, we report on a comparative study of (a) a standard paper prototype of a mobile application used to perform an early-phase seated (static) usability evaluation, and (b) a pseudo-paper prototype *created from the paper prototype* used to perform an early-phase, *contextually-relevant*, *mobile* usability evaluation. We draw some initial conclusions regarding whether it is worth the added effort of conducting a usability evaluation of a pseudo-paper prototype in a contextually-relevant setting during early-phase user interface development.

Keywords: mobile technology, paper prototyping, mobile lab-based usability evaluations.

1 Introduction

The benefits of paper prototyping are well recognized for desktop system design. Virzi *et al.* [1] compared the number of usability problems found via a think-aloud protocol using a low-fidelity prototype (a paper prototype) to the number found using a high-fidelity prototype (a final product) of an electronic encyclopedia. They found relatively little difference with respect to the number of usability problems uncovered using the two different prototypes; their findings suggest that a low-fidelity prototype is just as capable of finding usability problems as a high-fidelity prototype at a comparable degree of sensitivity. In a similar study, Sefelin *et al.* [2] compared a low-fidelity *paper* prototype to a low-fidelity *computer-based* prototype in order to determine whether participants in a usability study would be more likely to critique one or the other. They found that the prototype medium was relatively unrelated to the willingness of participants to make critical suggestions about the system design.

Less is known, however, about the benefits of *paper* prototyping for evaluation of *mobile* application designs [3]. Mobile systems are typically used by people who are *mobile*, in dynamically changing, contextually-rich and complex environments; many of the usability problems within mobile application designs are, therefore, best discovered through evaluation of the system in environments representative of the real world [e.g., 4, 5, 6]. Although mobile system design could potentially benefit from early-stage usability studies based on paper prototypes, such studies are rarely performed due to challenges presented by the *mobile* use of such prototypes [3].

Hendry *et al.* [7] created a paper prototype of a mobile application using a cardboard box the size of a table PC and laminated cards to simulate the screens of the user interface. To test their design with their target users, Hendry *et al*. conducted an 'on the street' field trial. Although they recognized the immense benefit in using early-stage prototypes with target users in situ, they reported great difficulty in using the prototype in the field: although participants were seated when using the prototype, and so held the box on their laps, they were forced to keep all the other prototype components on the ground or in pockets and, as a result, they found that the prevailing wind was a particular nuisance.

Sá and Carriço [3] discuss their experiences with low-fidelity prototypes for two different mobile applications. They initially constructed their prototypes out of card and Post-It notes, but observed that the dimension, weight, and handling of these truly paper prototypes "misled" participants about the form factor of the final products. They also found their paper prototypes to be too fragile; they weakened in structural integrity when used in the same manner as a final product (e.g., when placed in pockets, etc.). To address these concerns, Sá and Carriço created wooden frames that approximated the size, shape, and configuration of the final devices; these were then used to hold "screen cards". They found that this solution not only proved durable, but it also allowed users to comment on the shape of the device and the placement of buttons relative to how they held the device.

The research presented in this paper represents an *initial* investigation into the comparative strengths of a traditional paper prototype used in a seated evaluation protocol to a pseudo-paper prototype (created from the paper prototype) used in a mobile, lab-based protocol, in terms of the number and severity of usability problems identified using each. This study draws on research in the field of *effective* mobile, *lab-based* usability evaluation [e.g., 6, 8, 9] to expand on the findings of Sá and Carriço [3] and Hendry *et al.* [7] in order to further our understanding of potential mechanisms by which to effectively (and conveniently) use paper prototypes in *lab-based, mobile* evaluations, as well as to discover the relative merits of doing so (as compared to simply employing a traditional seated protocol). The following sections describe our evaluation design and process, and discuss our results, respectively. We conclude, in Section 4, with a brief discussion of further work.

2 Evaluation Design and Process

Our study was based on a paper prototype of a mobile system designed to be used in a grocery store to enhance the shopping experience. It was, in essence, a shopping cart-mounted, shop-and-scan system designed to support consumers' choice of products based on health attributes, price, or customer ratings. Figure 1 shows a sample screen from the paper prototype.

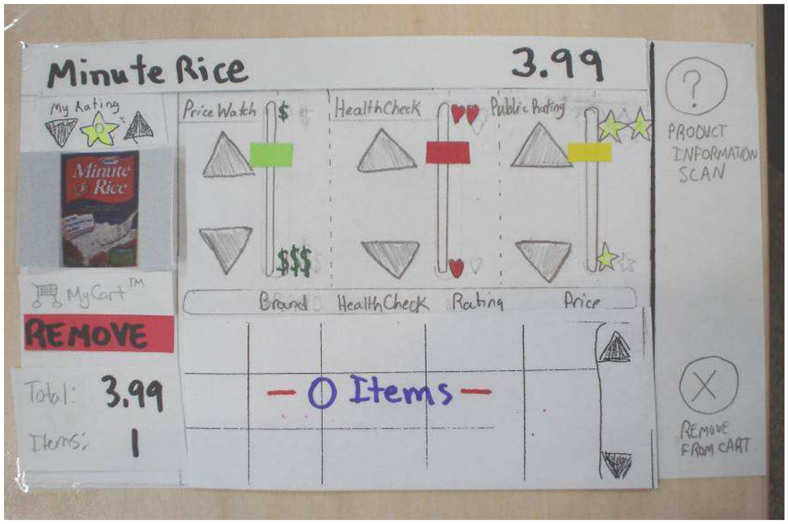

Fig. 1. Sample screen from the paper prototype.

The dimensions of the paper prototype reflect the actual screen size of the end device (a Fujitsu tablet) so as to prevent misleading users on issues of screen dimension. Specific screen mock-ups were developed based on two shopping scenarios which we used in our study.

Taking as inspiration a tool developed by Sá and Carriço [3], we developed a *pseudo*-paper prototype of our application based on the paper prototype shown in Figure 1. We took a series of digital photographs of the paper prototype at each stage during a walkthrough of our two shopping scenarios; we then organized these photographs into a PowerPoint presentation, linking them together by creating invisible clickable areas over the prototype's buttons – thus allowing participants to progress through the various screens associated with our study scenarios. We installed our pseudo-paper prototype on our Fujitsu tablet; thus participants could interact with the clickable areas by tapping the touchscreen of the device. To handle system response where a user 'scanned' a product (i.e., where interaction would be focused on a scanner rather than the touchscreen), we developed a secondary wireless application to allow us to remotely advance the PowerPoint slide when we saw a user 'scanning' a product.

For manageability, we did not 'activate' components of the photographs that would have constituted incorrect user actions relative to our scenarios; during our evaluations, we were able to observe such action intentions, but our pseudo-prototype limited users' ability to follow through on such intentions. Similarly, after piloting the use of the paper prototype (with the complexity of individual paper components for each screen) we decided to work with printouts of the same digital photographs used for our pseudo-paper prototype, and to verbally limit activation of 'misguided' user intentions. Although we recognize the limitations of this approach in terms of curtailing exploratory behavior and preventing, to some extent, observation of recovery behavior, we feel our decision placed both prototypes on an even basis for the purpose of our study – namely, to compare the impact of the *evaluation protocols* rather than conduct an in-depth evaluation of our specific design; furthermore, we eliminated, for the paper prototype, some of the lag in 'screen updates' brought about by the need for the evaluator to manually reset the prototype after each user action, thereby bringing it more in line with the pseudo-paper prototype in this regard.

Appropriately designed lab-based studies have proven a viable means by which to meaningfully assess the usability of mobile applications under controlled, experimental conditions [8, 9]. For the purpose of evaluating our pseudo-paper prototype, we therefore designed our study to reflect (albeit, abstractly) realistic environmental conditions – namely, a grocery store.

Fig. 2. Participant using the pseudo-paper prototype in contextually-rich, lab-based study.

Figure 2 shows our experimental set-up for our pseudo-paper prototype. We mounted the tablet onto a 'shopping cart' which participants were required to navigate around a series of 'aisles' in order to select and scan products (props that were attached to our 'aisles') based on provided shopping lists/scenarios. The evaluator followed each participant in order to determine when a product was scanned, and to subsequently progress the PowerPoint slide as applicable. As participants were completing their study tasks, they were surrounded by ambient grocery store sounds [10] played at ~69dB(A) (based on real world readings we had taken previously).

Our study design for the paper prototype simply required participants to interact with the paper version of the prototype whilst seated at a desk; the evaluator 'acted' as the computer, manipulating the components of the prototype in response to participants' actions. To bring parity to the two studies, we used the same ambient grocery store noise in this study set-up; additionally, we provided participants with the product props for the products itemized in the shopping lists/scenarios, albeit they were just available on the table next to the participants.

We adopted a between-groups design, assigning participants to one of two groups based on prototype. Participants in each group were given minimal training in the use of the system since it was designed to be 'walk-up-and-usable' without training. All participants were required to work through the same two provided shopping scenarios, the order of completion being counterbalanced across participants in each group. Across both groups, we used a think-aloud protocol combined with audio/video recordings of users' commentary and actions. Twelve people participated in our study, six per prototype/group.

3 Results and Discussion

We generated a content log of participants' activities and commentaries based on the audio/video recordings for each session. We applied two qualitative analysis techniques to the content logs. In the first instance, we conducted a usability defect analysis to compare the types and distribution of usability defects identified using the two prototypes; we then performed a heuristic analysis based on the content log data to determine the uniqueness and severity of problems found using each prototype. The following sections reflect on our analyses.

3.1 Usability Defect Analysis

Each content log was analyzed to identify and tag usability problems according to Lindgaard's [11] categorization of usability defects. Figure 3 lists the defect categories, and shows the number of instances of each according to prototype.

In *general*, the distribution of usability defects follows a similar pattern across the two prototypes. Two exceptions to this lie in the *Screen Design & Layout* and *Terminology* categories; while, for the paper prototype, we see a drop in the number of usability defects in these two categories compared to the *Navigation* category, we see the opposite for the pseudo-paper prototype. Both the *Screen Design & Layout* and *Terminology* categories relate directly to how quickly users can find and recognize user interface elements. A possible explanation for the divergence across the two prototypes for these categories could be that, on account of the fact that they were seated and not required to multitask, participants using the paper prototype may have benefited from an increased capacity, or felt it more appropriate, to scrutinize these aspects across the whole design; conversely, on account of the fact that they were required to multitask, participants using the pseudo-paper prototype likely only paid direct attention to *specific* aspects of the layout and terminology of the interface as/when they were needed. Additionally, given the contextual relevance of the study protocol, participants using the pseudo-paper prototype may have felt it contextually inappropriate to clinically 'examine' the design, but instead felt compelled to simply use the system much as they would in the real world. This suggests that the results from the pseudo-paper prototype might better reflect the realistic 'walk-up-and-usability' of the design.

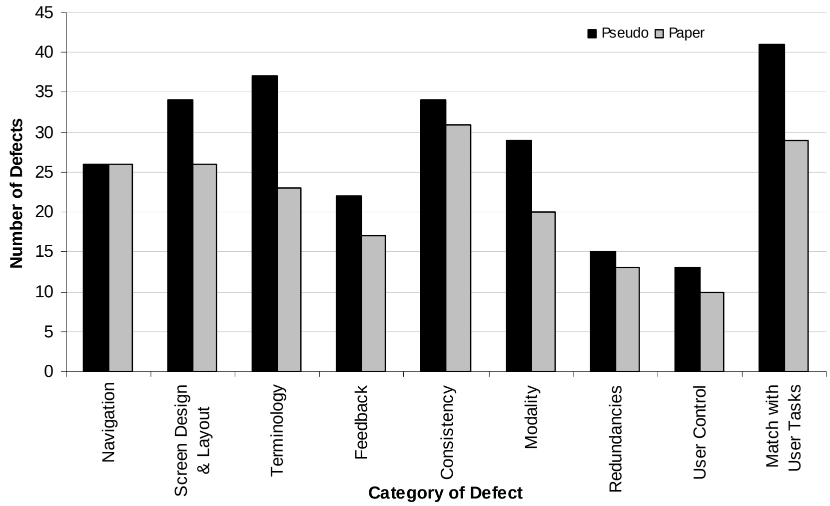

Fig. 3. Total number of usability issues identified according to category and prototype.

Across all defect categories, the number of instances identified by participants using the pseudo-paper prototype is equal to or (often substantially) greater than the number identified by participants using the paper prototype (see Figure 3); in total, 251 defects were identified using the former, compared to 195 identified using the latter). This difference is particularly noticeable in the number of *Terminology* and *Match with User Tasks* defects identified using each prototype. We suggest that the substantial increase in number of defects identified in these categories by participants using the pseudo-paper prototype demonstrates the potential impact of the contextually-relevant, mobile setting in which the prototype was evaluated. That is, when participants were required to interact with the system in a setting that (a) reflected the need to actively multitask and (b) the cognitive challenges associated with shopping in the real world, the inappropriateness of terminological aspects of the design, as well as the limited degree to which the design fitted with the primary tasks of the user, became obvious; the use of the paper prototype did not support such extensive observations. We would suggest that the use of the pseudo-paper prototype within a contextually-relevant, mobile protocol has proven noticeably more effective at enabling us to identify usability defects in the design, and to highlight where the bulk of defects lie with respect to realistic usage scenarios.

3.2 Heuristic Analysis

Each content log was again analyzed, this time to identify and tag usability problems according to the mobile heuristics appropriated by Bertini *et al*. [4], as summarized in Table 1. Like most usability heuristics, Bertini *et al's* mobile heuristics are intended to be used by usability experts during direct, hands-on analysis of a user interface design. By applying the heuristics to the *observed* interactions of *test users*, we appreciate that our use of the heuristics is slightly unorthodox, but we simply used them to provide an alternative means by which to classify the usability problems we observed. Given the typical complexity of context of use of mobile applications (as highlighted in our case study) it is imperative that the heuristics proposed by Bertini *et al* are observed if true mobile usability is to be achieved; as such, we felt these heuristics provided a good comparative measure of the efficacy of our evaluation protocols.

Table 1. Summary of mobile usability heuristics.

Heuristic	Description
1	Visibility of system status & losabilty/findability of the mobile device
2	Match between system and the real world
3	Consistency and mapping
4	Good ergonomics & minimalist design
5	Ease of input, screen readability, and glancability
6	Flexibility, efficiency of use, and personalization
7	Aesthetic, privacy, and social conventions
8	Realistic error management

Figure 4 shows the distribution (or nature) of *unique* usability problems found using each of the prototypes, as well as the unique instances of common (overlapping) usability problems found using both. Overall, participants using the pseudo-paper prototype found the most unique usability problems (27); in contrast, participants using the paper prototype found 19 unique problems, with an additional 19 problems being found irrespective of prototype.

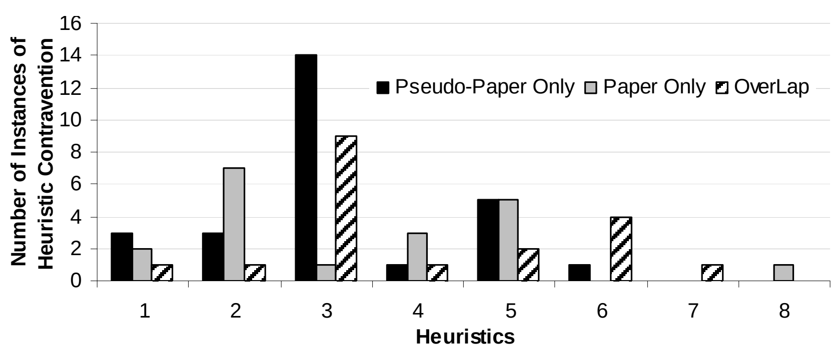

Fig. 4. Distribution of unique usability problems according to heuristic and prototype (including problems identified using both prototypes – overlap).

As can be seen from Figure 4, there were more instances of mismatch between the user interface design and real world context (Heuristic 2) identified using the paper prototype than the pseudo-paper prototype. We suggest (on the basis of user comments and our observations) that this was because, by using the pseudo-paper prototype in a contextually-relevant, mobile setting, participants were better able to match aspects of the system to the context in which the system was designed to be used; conversely, seated participants using the paper prototype were forced to 'imagine' the context of use, and as such often over, or inappropriately, analyzed the situation.

Although there was considerable overlap in the defects found according to Heuristic 3 (*consistency and mapping*), there was also a large gap in the number of such usability problems identified via the use of the pseudo-paper prototype compared to the paper prototype. The focus of Heuristic 3 is closely related to, and concurs with the findings for, the *Screen Design & Layout*, *Terminology*, and *Match with User Tasks* defect categories discussed previously: consistency and mapping are concerned with how participants *think* things should work, and this internalized mapping is based, in large part, on the clarity of screen design, including terminology used.

Neither prototype led to identification of many *aesthetic, privacy, and social convention* problems (Heuristic 7); this is unsurprising, given that both prototypes were being used in a lab without other people around. Whilst there is scope to remedy this in the mobile, contextually-relevant evaluation protocol supported by the pseudo-paper prototype, it is doubtful this could be addressed using the paper prototype.

Each identified and classified usability problem was additionally given a severity rating based on Nielsen's Severity Rating Scale, as used by Bertini *et al.* [4], namely: cosmetic; minor; major; and catastrophic. Figure 4 shows the breakdown of severity of the unique, heuristically-derived usability problems according to prototype.

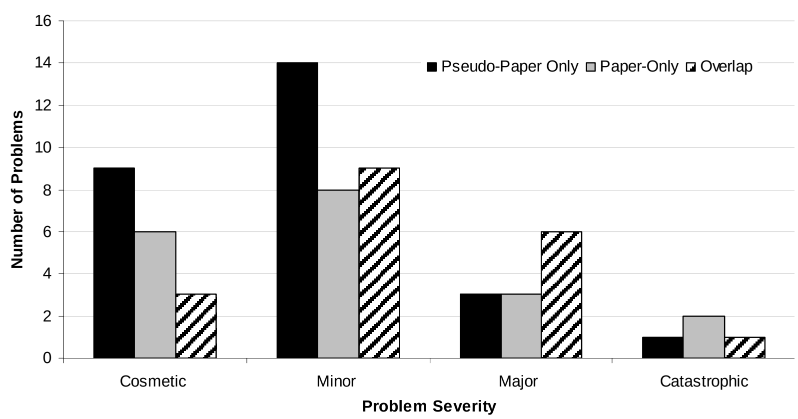

Fig. 5. Severity of problems according to prototype.

Encouragingly, the use of the pseudo-paper prototype not only identified more usability problems, but the problems identified were distributed across all 4 levels of severity; with the exception of *catastrophic* problems, the relative distribution of problem severity was similar for both prototypes. Participants using the paper prototype identified two unique catastrophic problems that were not noted by participants using the pseudo-paper prototype; on closer inspection of the specifics of these problems, however, we noted that they related to knowing how to 'check-out' and handle fruit and vegetables – neither of which were part of the evaluation scenario, and neither of which were a noted issue when the contextual-relevance of the pseudo-paper prototype environment was present. As a consequence, in our case, we do not see the lower extent of identification of catastrophic problems when using the pseudo-paper prototype to be a drawback of the prototype, but rather a reflection of its ability to mediate the severity of problems relative to context.

4 Conclusions and Further Work

We attribute no statistical significance to the observations we have presented in this paper; instead, we present our results as an *initial observation* of the differences in usability problems identified using a paper prototype of a mobile application in a traditional static evaluation setting versus a pseudo-paper prototype of the same application in a contextually-relevant mobile evaluation setting (that is, taking advantage of the contextual relevance the pseudo format could support in its respective evaluation protocol). We have shown that the use of the pseudo-paper prototype allows participants to identify more usability problems (whilst maintaining a similar distribution of defects to that observed with the paper prototype), to identify more unique usability problems (again, preserving distribution), and to be compatible with the paper prototype in terms of supporting the identification of usability problems across the various severity levels. We have also shown the benefits of a pseudo-paper prototype in terms of its ability to be used within a contextually-relevant experimental protocol, such that the problems identified better reflect what might happen in real use – that is, we consider the results more *meaningful*.

We have, obviously, only compared the paper and pseudo-paper prototypes relative to one application domain. We feel it would be beneficial to repeat the evaluation for additional application domains to determine the generalizability of our findings. Once in possession of such data, we would then be in a position to conduct deeper, statistical analysis to further demonstrate the merit of conducting *contextually-rich*, *mobile* usability evaluations of *pseudo*-paper prototypes in the early-phases of mobile UI design. We also anticipate comparing the benefits of our *pseudo*-paper prototyping approach to other, increasingly established, mechanisms for evaluation of mobile user interface designs. To conclude, therefore, we present, here, the results of our existing qualitative analysis as a *first* indication of the potential usefulness of such an approach in the hope that developers can begin to benefit from this early investigation.

References

1. Virzi, R.A., Sokolov, J.L., and Karis, D., Usability Problem Identification Using Both Low- and High-Fidelity Prototypes, In: Proceedings of SIGHCI Conference on Human Factors in Computing Systems (CHI'96), pp. 236-243, Vancouver, Canada, April 13-18, (1996).

2. Sefelin, R., Tscheligi, M., and Giller, V., Paper Prototyping - What is it Good For? A Comparison of Paper- and Computer-Based Low-Fidelity Prototyping, In: Proceedings of Conference on Human Factors in Computing Systems - Extended Abstracts (CHI '03), pp. 778-779, Ft. Lauderdale, USA, April 5-10, (2003).
3. Sá, M. and Carriço, L., Low-Fi Prototyping for Mobile Devices, In: Proceedings of SIGCHI Conference on Human Factors in Computing Systems (CHI '06), pp. 694-699, April 22-27, (2006).
4. Bertini, E., Gabrielli, S., and Kimani, S., Appropriating and Assessing Heuristics for Mobile Computing, In: Proceedings of Workshop on Advised Visual Interfaces (AVI'06), pp. 119-126, Venezia, Italy, May 23-26, (2006).
5. Kjeldskov, J., Graham, C., Pedell, S., Vetere, F., Howards, S., Balbo, S., and Davies, J., Evaluating the Usability of a Mobile Guide: The Influence of Location, Participants, and Resources, Behaviour and Information Technology, 24(1), pp. 51-65, (2005).
6. Kjeldskov, J. and Stage, J., New Techniques for Usability Evaluation of Mobile Systems, International Journal of Human Computer Studies (IJHCS), 60(5-6), pp. 599 - 620, (2004).
7. Hendry, D.G., Mackenzie, S., Kurth, A., Spielberg, F., and Larkin, J., Evaluating Paper Prototypes on the Street, In: Proceedings of Conference on Human Factors in Computing Systems - Extended Abstracts (CHI '05), pp. 1447-145, Portland, USA, April 2-7, (2005).
8. Kjeldskov, J., Skov, M.B., Als, B.S., and Høegh, R.T., Is It Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field, In: Proceedings of 6th International Symposium on Mobile Human-Computer Interaction (MobileHCI'04), pp. 61 - 73, Glasgow, Scotland, September 13 - 16, (2004).
9. Lumsden, J., Kondratova, I., and Langton, N., Bringing A Construction Site Into The Lab: A Context-Relevant Lab-Based Evaluation Of A Multimodal Mobile Application, In: Proceedings of 1st International Workshop on Multimodal and Pervasive Services (MAPS'2006), pp. 62 - 68, Lyon, France, June 29, (2006).
10. PacDV, Grocery Store, http://www.pacdv.com/sounds/ambience_sounds.html.
11. Lindgaard, G., Usability Testing & System Evaluation: A Guide for Designing Useful Computer Systems, 1st ed. London: Chapman & Hall, (1994).