05-410/610 Intro to HCI Methods Fall 05

Homework 7

Retrospective

Group E

Colleen Koranda

Hwi Kyoung Lee

Jack Li (Manager)

Prudence Marasigan

Samantha Konwinski

December 7, 2005

Retrospective

Overview

To generate and validate our Palm Pilot redesign ideas, we employed several design and usability analysis methods, including Contextual Inquiry (CI) and Contextual Design (CD), Keystroke-Level Modeling (KLM), Heuristic Evaluation (HE), Cognitive Walkthrough (CW), and Think-Aloud (TA). The following is a retrospective account of our experiences and reflections with these methods.

Contextual Inquiry

CI provides the foundation for user-centered design allowing us to gather objective data in the early stages of exploration. It provides qualitative data about a user's environment, roles, tasks, interactions, culture, and breakdowns through observation and conversations in the context of her work. This rich field data gave us insights on the design that would best fit the user's work demands. CI allows probing into the user's activities from different perspectives to understand the user's needs including ones she herself may not be able to identify.

Having only one user participate in CI, we may have missed out on greater insight on other ways to-dos are managed by others. Additional CIs would have helped focus our ideas early on. Another challenge was that we did not observe a user interact with a Palm during the CI. We would like to have to have performed an additional CI with a Palm user to have more data directly related to Palm usage, to get an understanding of

the needs of that type of user. Thus, we had to keep in mind this type of user and her activities as we explored the capabilities of the Palm and applied the usability methods.

In comparison to the other methods, CI (along with CD) provides the largest amount of raw data and is most appropriate in the early stages of design exploration. KLM, CW, and TA are task-specific methods while CI covered the holistic work practice. Although HE easily addresses a wider range of tasks, it fails to tie in the target user and that user's context, which CI and CD cover. TA also generates grounded objective empirical data like CI and CD, but as mentioned before does not cover the context of the user's tasks.

The main barriers to the CI's validity come directly with how the interview was performed. For instance, asking leading questions would bias the user's response and the data gotten; as a result, expertise in conducting a well-focused CI is crucial. Another cost is, besides having expert interviewers, the CI users themselves, who often need compensation. CIs themselves are time-consuming, magnified by the fact that multiple CIs often need to be conducted because having one user may not be representative of the population.

Contextual Design

CD creates models from CI data to provide an environment for brainstorming design concepts that fit into the user's work practice in this early exploration phase. Creating

models as a group ensured that the entire team had a shared understanding of the data, and grounded the beginning of our design process. The different visual representations focus the team into the various faces of the work. Flow captures information communication, Cultural captures invisible influences, Artifact captures information organization, Sequence captures task structure, and Physical captures environment layout.

We found it challenging that the interviewer was not present during our interpretation session. At one point in the CI, the interviewer made a comment about “good buckets” (CI L030). Situations like these made it difficult to get a full understanding of the entire interview, especially when we did not understand what the interviewer was trying to probe. We had to use our best judgment by reviewing the CI video multiple times. It was extremely helpful to keep track of these questions along with design ideas and insights as extra triggers to our mindset.

The Flow model was especially challenging because we had to decide the level of detail to include. By having only one CI to review, we did not gain from the benefits of consolidation, and it was hard to see what data would be most influential for a redesign. An additional limiting factor was the scope of the homework for the Sequence and Cultural models. By modeling just a section of each, these models could have generated even more design ideas for us.

CD gave us a detailed view into the various aspects of a busy user, given the numerous tasks she has to manage. Her many activities led us to focus on a particular task from which to generate design ideas for the Palm. For example, our Artifact and Cultural models express the importance of having items visible and in her face. The Flow revealed the high work load and scattered nature of keeping track of her to-dos.

KLM

GOMS is a technique that makes a priori predictions of performance time of a benchmark task involving routine cognitive skill. It provides a qualitative explanation of why specific keystroke-level operators of the task require a certain amount of time. KLM is the simplest GOMS technique, which has software support in CogTool that allows designers to use human data provided by researchers to make these predictions. CogTool allows designers to compare different redesign ideas of the same task.

KLM allowed us to understand where inefficiencies in design could creep in. Having the software support of CogTool significantly reduced the resources needed to obtain such useful data. Even though the homework task did not directly relate to our redesign, we found KLM to be useful in understanding how widget changes can affect task time. For example, increasing the size of a button could reduce task time. KLM ensures in a quantitative way that our design ideas will hold beyond just novice uses.

In general, KLM provides the most undisputable information because its results are objective and quantitative in its output of a calculated time. Performing KLM requires

an interface to be generated so it is more applicable later in the evaluation stages of design, just like the other usability evaluation methods HE, CW, and TA. HE also uses quantitative figures in its analysis although they are in the form of ratings justified subjectively whereas KLM's time predictions come from objective empirical research. KLM, CW, and TA are all very task specific methods; unlike CW and TA, KLM does not consider a specific user during analysis. With CW, although the user is a fictional character derived of assumptions made by the design team, considerations are made about whether the user will perceive widgets or understand meanings. With TA, we were able to directly see how a real user reacted to a specific task. KLM is meant to evaluate an expert user's completion time for a task with no errors.

The main shortcomings of KLM are that it cannot predict complex mental tasks and that it does not deal with standard cognitive issues such as how readable words are and how recognizable icons are. KLM does not address many critical areas of the Palm when compared with HE because it is so task specific. Thus, the principles of KLM are useful in identifying small inefficiencies like button think times, but do not identify bigger goal-related problems. Although CogTool makes a time calculation of the task, it would have been helpful if more details about specific subtask times were given to further analyze bottlenecks of the task and to see how keystroke-level operators compared.

Heuristic Evaluation

HE occurs in the evaluation stages of the design process, using ten heuristics to guide an interface inspection. It can be used anytime some form of an interface has been designed. The cost to perform HE is significantly lower because of the amount of problems that can be identified with few people; however, its effectiveness in finding problems increases with the number of evaluators and their expertise levels. The severity ratings allow problems to be assigned priority in a seemingly quantitative manner because each evaluator can have a different opinion on how important a problem is.

HE allowed our design team to dive into many Palm interface problems. Although four members of the evaluation team had never used a Palm, we were still able to identify potential usability problems by comparing the interface against the given heuristics at the second pass after the first pass-around to explore the interface. The ten heuristics allowed us to look for a broad range of problems within a common ground.

Problems identified during HE were valuable for analysis. We used the results as a motivational resource to generate design ideas and a problem list for small fixes. This method identified many usability problems, some that first appeared to be mainly cosmetic. However, after consolidating the data, we were able to predict several significant problems with the interface, and a rough estimate of other high-level problems. We identified problems that were useful justifications in our redesign; for example we found that an event with an un-named subject would not be saved, there

are limitations in sorting to-dos, and there is no quick way to jump back to the calendar application after viewing to-do items.

HE is known to produce the most problems with smallest cost, when compared to other methods. In our experience, we generated over forty problems of varying importance. This shows that HE is unreliable for solely generating critical problems, because of the subjectivity of individual evaluators. HE allowed us to look at a wide range of problems, but the focus of some items was outside of the scope of the task we addressed in our redesign based on our CI data, such as the Private functionality. These are examples of threats to validity within HE.

Data gathered during HE complemented our usability analysis quite well. At first, we were hesitant to resolve usability issues that had solely arisen from HE because of the threats to validity. However, after reviewing TA data, HE data helped us define the problems the user encountered more precisely. For example, we detected that the unnamed event that the user in the TA study had set would be deleted since we had previously identified it in HE. HE allowed us to see the problem space in a broad view; HE data greatly assisted observed data in TA to substantiate persuasive usability issues. Having done an HE before CW and TA, we were able to use HE to help validate problems found in these other methods. Although it is not a user-centric method, HE provided insights that can be used hand-in-hand with the other methods to confirm problems and recommend solutions.

Cognitive Walkthrough

CW is a usability inspection method used to evaluate whether a user will be able to complete a benchmark task while exploring the interface. It precisely locates where a user might fail and where problematic interface elements are, because the method exhausts four criteria in each state relevant to the task process. CW occurs in the evaluation phase to identify how a user would interact with an interface.

Through the CW process, we concentrated on the understanding that users are not like us. Because of the possibility that different user assumptions would lead to either a more failure or success-based story, it was important to pay attention to what we assumed about the user's technical abilities and understanding of the Palm.

We initially struggled to differentiate between the perception and comprehension criteria. However, as soon we saw the separation between them, we took advantage of the split to pinpoint at what level of mental process there will be a failure and what is the best design idea to address these failures. The failure stories in perception criteria are influenced by the visual representation of the system. Improvement in the layout, such as, more prominent widgets and positioning of widgets, was the common factor in the design ideas that we developed from these failure stories. On the other hand, failure stories in comprehension are dependent on the knowledge and experience of the user. Another benefit of CW is the feedback criterion. We realized that it is crucial to inform the users if they are in the right track or not. This feedback informs the user not only the transition of state but also the confirmation for the previous action.

Like KLM and TA, CW is very task specific. CW differs from these other methods because a specific user is defined for the analysis. A priori assumptions are made and these impact the results of the method. If too many or too few assumptions are made, valuable data may not be generated.

KLM, CW, and TA are all very task specific methods; unlike CW and TA, KLM does not consider a specific user during analysis. With CW, although the user is a fictional character derived of assumptions made by the design team, considerations are made about whether the user will perceive widgets or understand meanings. With TA, we were able to directly see how a real user reacted to a specific task.

We found only five failures using CW, but found more than forty using HE. We felt that CW was not as thorough as some of the other methods. Much of the reason for this is that we were not limited to a specific calendar task using HE; we were able to review all possibilities related not just to the calendar, but also the task list program. In addition, we feel that CW is less useful when evaluating a current redesign because to an extent other methods like HE and TA can find similar problems. In our redesign, we only applied two of the five failures we identified. However, CW is very helpful in the redesign process so we can understand the mindset of a user and predict whether he will be successful or not in his goals.

Think-Aloud

TA, used in the evaluation phase of design, provides qualitative, subjective data in a laboratory setting by testing an interface with a user unfamiliar with it. Good aspects make great success stories and allow designers to keep a record of what already works well with the system to avoid redesigning what already works. It gives insight into what a user's goals really are, as well as the system response to user's actions.

TA was useful in justifying any misconceptions of the user we had in HE and CW. We found that we may have overestimated our assumptions of the a priori of our CW user, as well as the ease of the task. We found that our CW user had only five failure stories, but in TA, many more problems were identified related to navigating to the correct date and understanding button labels and screen purposes. Because of this, we were able to get good insight into problems users might have that we did not consider in previous methods.

While TA was able to demonstrate where a user may have the most difficulty and how emotionally taxing the current design may be, it was limited by the task demonstrated. We might have been able to identify more problems that were not classified as critical incidents but still may have supported our redesign ideas. Although our goal was to redesign to rectify the most serious usability problems, also recognizing the weaker problems may have provided more insight into a more holistic approach to the redesign. We also noted that conducting the user test on the Palm emulator made Palm-specific problems harder to identify, especially for members of our group who

are not familiar with the Palm. Like CI and CD, TA requires users, which may have added cost.

TA helped us sort out what problems really should be addressed in our redesign. The empirical nature of the method strengthened the validity of the problems we identified in other methods like HE and CW. It was also a useful technique to apply to our redesign because when we questioned how a user would react to our redesign, an outside view of our design showed that the user could understand the interface.

Conclusion

The reflection on each of the methods shows how they can be applied to aid the user in performing the task and improving the system to support these goals. Each method is unique and addresses a different way to understand the user, improve task performance, or improve usability. We learned a lot about understanding where problems may appear that are not obvious, and how to determine which problems are most critical to resolve. Our redesign ideas were grounded in the data that we found from these methods, and ensured that our changes will be beneficial for the user. The skills we gained from this experience will apply in our HCI project and in our future projects outside the university.