Experimenting on the Cognitive Walkthrough with Users

Wallace Lira  
Vale Institute of Technology  
Av. Boa Ventura da Silva, 400  
C.P. 668 – 13.560-970  
Belém, Brazil  
wallace.lira@gmail.com

Cleidson de Souza  
Vale Institute of Technology  
Av. Boa Ventura da Silva, 400  
C.P. 668 – 13.560-970  
Belém, Brazil  
cleidson.desouza@acm.org

Renato Ferreira  
Universidade Federal do Pará  
Rua Augusto Corrêa, 01  
C.P. 479 – 66075-110  
Belém, Brazil  
renpina@gmail.com

Schubert R Carvalho  
Vale Institute of Technology  
Av. Boa Ventura da Silva, 400  
C.P. 668 – 13.560-970  
Belém, Brazil  
schubert.carvalho@itv.org

Abstract  
This paper presents a case study aiming to investigate which variant of the Think-Aloud Protocol (i.e., the Concurrent Think-Aloud and the Retrospective Think-Aloud) better integrates with the Cognitive Walkthrough with Users. To this end we performed a case study that involved twelve users and one usability evaluator. Usability problems uncovered by each method were evaluated to help us understand the strengths and weaknesses of the studied usability testing methods. The results suggest that 1) the Cognitive Walkthrough with Users integrates equally well with both the Think-Aloud Protocol variants; 2) the Retrospective Think-Aloud find more usability problems and 3) the Concurrent Think-Aloud is slightly faster to perform and was more cost effective. However, this is only one case study, and further research is needed to verify if the results are actually statistically significant.

Author Keywords  
Usability Evaluation; Cognitive Walkthrough; Think Aloud-Protocol; Mobile apps.

ACM Classification Keywords  
H.5.2. [User Interfaces]: Miscellaneous
**Introduction**

Usability can be defined as “a fairly broad concept that basically refers to how easy it is for users to learn a system, how efficiently they can use it once they have learned it, and how pleasant it is to use” [13]. The usability of mobile applications has been of great concern to the industry. For example, Apple and Google both have published user interface guidelines regarding mobile apps [2, 4]. Due to these reasons, it is necessary to apply usability tests to evaluate the usability of mobile applications. Popular usability testing methods include both the Think-Aloud Protocol [9] and the Cognitive Walkthrough [13].

The Think-Aloud Protocol instructs the evaluators to observe and collect feedback from a user interacting with the evaluated software. The user should speak everything he is thinking of regarding the evaluated software. This work focuses on two variants of this method: the Concurrent Think-Aloud, which instructs the user to talk while he is interacting with the software; and the Retrospective Think-Aloud that instructs the user to talk while he watches a video of his interaction with the software [9].

The Cognitive Walkthrough, on the other hand, provides a test process which includes four heuristics to be used in the inspection of each possible user action by an usability expert. There are some variants of the Cognitive Walkthrough, which were compiled in a paper by Mahatody et. al. [8]. The Cognitive Walkthrough with Users [5] was one of them. This usability testing method blends both the Concurrent Think-Aloud and the Cognitive Walkthrough. However, Mahatody et. al. [8] wonders whether the Concurrent Think-Aloud is the most appropriate usability testing method to be merged with the Cognitive Walkthrough with Users. The objective of this work is to investigate which variation of the Think-Aloud Protocol (Concurrent or Retrospective Think-Aloud) integrates best with the Cognitive Walkthrough with Users. The evaluated application is simply referred as Digital Plants, which is an app for automatic plant identification and classification using leaf images.

This work is organized as follows. The section Case Study details of its execution. The findings related to the experiment conducted can be found in the Findings section. The last section presents the conclusions and some possible future works that could follow up this research.

**Case Study**

This work presents a case study to investigate if there are differences in the results of the Cognitive Walkthrough with Users should the variation of the Think-Aloud Protocol (Concurrent or Retrospective Think-Aloud) integrated with it change. This was referenced as an interesting course of action [8]. Furthermore, the two variants of the Think-Aloud Protocol can achieve different results [16].

**Evaluated Application**

The use of leaf images for automating plant identification and classification processes has been the subject of numerous studies [1, 6, 15]. The main features extracted from leaf images are based on their shape and venation. Leaf venation will be referred as the leaf texture when describing the Digital Plants’ functions even though it is possible to extract other data by analysing the leaf’s texture.
The main goal of the Digital Plants app is to automate the plant identification process. The app should also provide a plant catalogue with extensive information on them.

The app’s intended use is to help botanists to identify the species of plants through a single leaf without harming the plant. Today, this process is performed manually with the help of textbooks and people without formal training but with vast experience regarding the local flora.

The Fig. 1 presents the Digital Plants functions in a mind map. There are two main modules on the tool: Plant List and Leaf Analysis. We applied principles of mobile usability [2, 12] at the development the Digital Plants app user interface. We can see an example of the tested user interface at the Fig. 2.

**Participants**

There were 12 users in the case study. It is widely accepted that only five users are enough to find around 85% of the usability problems using the Think-Aloud Protocol [11]. Therefore 6 users were subjected to each variant of the Think-Aloud Protocol, which gave us enough information for comparing both usability evaluation methods. Table 1 shows the age, which mobile operating systems he had experience on, for how long he owns a mobile device with that operating systems and to what variant of the Think-Aloud Protocol he was subjected to.

The mean participant age in years is 33.25 and the standard deviation is 7.14. The mean experience with mobile operating systems in years is 1.01 while the standard deviation is 0.9. Furthermore, each of the participants are researchers working at Vale Institute of Technology. The Think-Aloud Protocol variant that each participant was subjected to was determined randomly.

**Table 1: Participant characteristics.**

<table>
<thead>
<tr>
<th>ID</th>
<th>Age</th>
<th>Mobile OS</th>
<th>How many years</th>
<th>TA Variant</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>29</td>
<td>Android</td>
<td>0.3</td>
<td>Concurrent</td>
</tr>
<tr>
<td>P2</td>
<td>40</td>
<td>iOS and Android</td>
<td>3</td>
<td>Concurrent</td>
</tr>
<tr>
<td>P3</td>
<td>35</td>
<td>-</td>
<td>0</td>
<td>Concurrent</td>
</tr>
<tr>
<td>P4</td>
<td>23</td>
<td>Android</td>
<td>1</td>
<td>Concurrent</td>
</tr>
<tr>
<td>P5</td>
<td>30</td>
<td>Android</td>
<td>1.4</td>
<td>Concurrent</td>
</tr>
<tr>
<td>P6</td>
<td>40</td>
<td>-</td>
<td>0</td>
<td>Concurrent</td>
</tr>
<tr>
<td>P7</td>
<td>48</td>
<td>-</td>
<td>0</td>
<td>Retrospective</td>
</tr>
<tr>
<td>P8</td>
<td>30</td>
<td>-</td>
<td>0</td>
<td>Retrospective</td>
</tr>
<tr>
<td>P9</td>
<td>24</td>
<td>Android</td>
<td>2</td>
<td>Retrospective</td>
</tr>
<tr>
<td>P10</td>
<td>32</td>
<td>Android</td>
<td>2</td>
<td>Retrospective</td>
</tr>
<tr>
<td>P11</td>
<td>37</td>
<td>Android</td>
<td>1.5</td>
<td>Retrospective</td>
</tr>
<tr>
<td>P12</td>
<td>31</td>
<td>Android</td>
<td>1</td>
<td>Retrospective</td>
</tr>
</tbody>
</table>

**Usability Testing Protocol**

The case study performed can be summarized by the following steps: Select Tasks, Select Participants, Perform the Traditional Cognitive Walkthrough, Perform the Think-Aloud Protocol Variant, Summarize the Usability Problems Found, Critical and Statistical Analysis.

In the first phase, we selected the tasks that were evaluated by the traditional Cognitive Walkthrough. Not every possible scenario must be evaluated, but we are recommended to select relevant ones. We selected 3 tasks: Discover where the plant with scientific name *Amelanchier Laevis* can be found; Discover to what species a real leaf belongs to through the leaf shape analysis; Discover to what species the same leaf belongs to through the leaf texture analysis.
The participants were selected and their characteristics (as presented in Table 1) were identified. The Cognitive Walkthrough was then performed. The usability experts identified usability problems that needed further investigation during the Cognitive Walkthrough. Additionally, the assumptions about the user experiences were based on the participant characteristics presented in the Participants subsection.

The next step aims at performing both Think-Aloud Protocol variants. In this step, the participants were randomly split in two groups. One of these was subjected to the Concurrent Think-Aloud and the other one to the Retrospective Think-Aloud. All the testing sections were recorded. We restricted the analysis presented in this paper to cognitive usability problems because it is possible to find them applying the Cognitive Walkthrough, so the matter of efficiency and satisfaction will not be addressed.

Afterwards we summarized the usability problems found by the Cognitive Walkthrough and by each version of the Think-Aloud Protocol. Finally, we should perform a critical analysis of the usability problems summarized. It is important to know if the problems that the usability experts found performing the Cognitive Walkthrough were really a concern to the users.

**Findings**

The number of usability problems of an user interface is $N(1 - (1 - L)^n)$ [11]. Where $N$ is the total number of usability problems in the user interface, $L$ is the proportion of usability problems discovered while testing with a single user and $n$ is the number of users that participated in the usability test. This formula means that around 85% of the usability problems should be found, assuming an $n$ value of 6, and an $L$ value of 31%.

We needed to evaluate whether the Retrospective Think-Aloud and the Concurrent Think-Aloud had similar results. This analysis turned out to reveal differences between the methods. We can observe from Fig. 3 that the Retrospective Think-Aloud found 66% more usability problems.
problems than the Concurrent Think-Aloud. The number of usability problems found performing both variants of the Think-Aloud Protocol seems to be converging to a maximum number of discoverable usability problems as the number of participants involved increases. This happens because the usability problems found by each user can also be found by the previous users, so we get diminishing returns when we increase the number of participants. This behavior is consistent with the results presented by [10].

Figure 3: Number of usability problems found and the number of participants involved.

We then sought to unveil whether the usability problems found by the Cognitive Walkthrough, Concurrent Think-Aloud and Retrospective Think-Aloud were similar. The Fig. 4 points out that the Retrospective Think-Aloud had comparatively identified more unique usability problems compared to the Concurrent Think-Aloud.

The high number of usability problems found only by the Cognitive Walkthrough method may imply that the evaluator was mistaken because the users did not judge the would-be problems a real issue. Even though this may happen, the Cognitive Walkthrough analysis can help the evaluator to improve his skills using that method because he had reliable feedback.

One of the possible reasons of the disparate result was that the participants felt generally more at ease when subjected to the Retrospective Think-Aloud as opposed to the Concurrent Think-Aloud. This may have happened because the Retrospective Think-Aloud allows the user to focus on the task at hand. However, the Concurrent Think-Aloud was 28% faster to perform with an average duration of 6.75 per user, with an associated standard deviation of 3.5, while these values for the Retrospective Think-Aloud were respectively 8.7 and 3.1. However, this data does not take into account the time that the usability evaluator took to analyse the footage of the users interaction with the application.

Finally, other metric that can be found analysing the data presented in this section is the time spent to find an usability problem. The Concurrent Think-Aloud was slightly more effective in this regard, finding 0.42 usability problems per minute. The Retrospective Think-Aloud, on the other hand, found 0.37 usability problems per minute.

Conclusions and Future Work
This paper presented a comparative analysis of how two Think-Aloud Protocol variants integrate with the Cognitive Walkthrough with Users. We found that both of these variants are equally compatible with the Cognitive Walkthrough with Users. Additionally, we found that the Retrospective Think-Aloud unveiled around 66% more usability problems compared to the Concurrent Think-Aloud. On the other hand, the Concurrent Think-Aloud was around 2 minutes faster to perform per user compared to the Retrospective Think-Aloud. Moreover, using the Concurrent Think-Aloud allowed the evaluator to find more usability problems per minute of users interaction, being slightly more cost effective compared to the Retrospective Think-Aloud.
It is important to note that the results presented in this work were obtained in only one case study. Therefore, to strengthen these conclusions we would need to run several sessions with different users and compare averages of usability issues found with each method associated with their margin of error.

**Acknowledgments**

We would like to thank the Vale Institute of Technology researchers for their participation in the usability tests.

**References**


