“Thinking About Thinking Aloud”: An Investigation of Think-Aloud Methods in Usability Testing

Obead Alhadreti
School of Computing Sciences
University of East Anglia
Norwich, NR4 7TJ, UK
College of Computer
Umm Al-Qura University
o.alhadreti@uea.ac.uk

Pam Mayhew
School of Computing Sciences
University of East Anglia
Norwich, NR4 7TJ, UK
p.mayhew@uea.ac.uk

Usability has become an imperative aspect of survival on the web, thus, it has always been considered as a crucial aspect of web design. This paper presents the results of a study that compared two think-aloud usability testing methods: the concurrent think-aloud and the retrospective think-aloud methods. Data from task performance, testing experience, and usability problems were collected from 40 participants equally distributed between the two think-aloud conditions. The results found that while the thinking aloud method had no impact on task performance and participants testing experience, participants using the concurrent think-aloud method detected a larger number of minor problems with the test interface than participants using the retrospective think-aloud method. These findings suggest a reason for preferring the concurrent think-aloud method to the retrospective one.

1. INTRODUCTION

When developing a software product, it is important to ensure that a high level of usability is attained. If products are not sufficiently usable, users will abandon them and find alternatives that better cater to their needs. As a result, effective usability evaluation methods are required to determine and improve the usability of software systems (Barnum, 2010). Over the last four decades, a number of different UEMs for determining the usability levels of software systems have been proposed. Among these methods, think-aloud (TA) methods, also known as TA protocols, are widely used (Van den Haak et al., 2009). According to Ericsson and Simon (1993), there are traditionally two basic types of TA methods: the concurrent think-aloud (CTA) method, in which participants think aloud as they carry out experimental tasks; and the retrospective think-aloud (RTA) method, where participants verbalise their thoughts after they have completed experimental tasks.

So far, the knowledge of contributions of TA methods to usability testing remains inconclusive, and research on usability testing methods has been criticised as being problematic and in a state of crisis (Hornbaek, 2010). Accordingly, the aim of this study is to investigate the use of the classic CTA method and the RTA method within laboratory-based usability testing.

2. APPROACH TAKEN

Given the research’s focus on investigating different variants of TA methods and the fact that TA testing methods are typically applied in usability laboratory settings (Norman and Panizzi, 2006), an experimental approach is used in the study. 40 university students were recruited for the experiment through purposive sampling and assigned to two groups, following between-subject design. The participants were asked to complete seven search tasks, and asked to fill in a post-test questionnaire. The two TA methods were compared through an evaluation of an online library website, which involved three points of comparison: participants’ task performance, test participants’ experiences, and number and type of usability problems discovered. The participants in the RTA condition were asked to watch their recorded performance on muted video and give retrospective reporting.
3. RESULTS

This section presents the results of the study in the following order: the participants’ task performance (subsection 3.1), the participants’ testing experience (subsection 3.2), and the usability problems detected (subsection 3.3).

3.1 Task performance

Table 1 illustrates the results derived from the measurements of the task performance. A Mann-Whitney test found no statistically significant difference in terms of the number of successful task completions and the time spent on tasks between the two TA conditions. This finding seems to lend support to (Ericsson and Simon, 1993) argument that thinking aloud does not have an effect on task performance.

<table>
<thead>
<tr>
<th>Performance measures</th>
<th>CTA</th>
<th>RTA (silence)</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task Completion</td>
<td>4.5</td>
<td>4.3</td>
<td>p = .499</td>
</tr>
<tr>
<td>Task Time (min)</td>
<td>20.6</td>
<td>18.9</td>
<td>p = .149</td>
</tr>
</tbody>
</table>

Table 1: TA methods and participants’ task performance

3.2 Participants testing experience

Table 2, 3 and 4 present the results of participants’ ratings regarding their testing experience. To begin with, all participants were asked to estimate in what respect(s) their working procedure on tasks differed from usual, by marking how much slower and more focused, they were while working on the tasks. As shown in Table 2, the participants in all two conditions felt that they had not worked all that differently from usual: the scores for the two items are rather neutral, ranking around the middle of the scale, and no significant differences were found between the conditions.

Table 2: TA methods and participants’ working condition

Participants were next asked how they felt about having to think aloud (concurrently or retrospectively) to which degree they thought this activity was difficult, unpleasant, tiring, unnatural, and time consuming. As is clear from Table 3, a Mann-Whitney test showed that there were no significant differences among the conditions regarding either of the questions. As the average scores of the participants ranged between 1.1 and 1.8, the participants clearly felt that they were not affected by the evaluator. In sum, the two usability test approaches revealed similar results with respect to the participants’ testing experiences. This finding is in agreement with (Van den Haak et al., 2009).

Table 3: TA methods and participants’ TA experience

3.3 Usability problems detected

Table 5 presents the mean number and standard deviation for problems detected per participant. It also classifies all problems according to their severity. A t-test test indicated that the participants in the CTA produced significantly more problems compared to the RTA participants, and found higher number of minor problems. These results echo that of (Peute et al, 2009).

Table 4: TA methods and the evaluator presence

<table>
<thead>
<tr>
<th>CTA</th>
<th>RTA</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unnatural</td>
<td>1.3</td>
<td>1.8</td>
</tr>
<tr>
<td>Disturbing</td>
<td>1.2</td>
<td>1.4</td>
</tr>
<tr>
<td>Unpleasant</td>
<td>1.1</td>
<td>1.3</td>
</tr>
</tbody>
</table>

Table 5: TA methods and usability problems detected

Eight-points scale (1: Strongly disagree to 5: Strongly agree)
3. REFERENCES


