Comparison of CTA and Textual Feedback in Usability Testing for Malaysian Users

Ashok Sivaji
MIMOS Berhad
Kuala Lumpur
Technology Park Malaysia
ashok.sivaji@mimos.my

Torkil Clemmensen
Department of IT Management
Copenhagen Business School
Solbjerg Plads 3
DK-2000 Frederiksberg
tc.itm@cbs.dk

Søren Feodor Nielsen
Department of Finance
Copenhagen Business School
Solbjerg Plads 3
DK-2000 Frederiksberg
sfn.mes@cbs.dk

Abstract

Usability moderators found that the concurrent think-aloud (CTA) method has some cultural limitation that impacts usability testing with Malaysian users. This gives rise to proposing a new method called textual feedback. The research question is to determine whether there are any differences in terms of usability defects found by employing the new method. Due to the high power distance, it is hypothesized that the CTA method may not be sufficient and hence a textual feedback method is recommended instead. Hence, the objective of this study is to determine if there are any differences in usability defects from the concurrent think-aloud (CTA) method (Condition 2) and textual feedback method (Condition 1) within the same group of Malaysian users. A pair-wise t-test was used, whereby users were subjected to performing usability task using both methods. Results reveal that we can reject the null hypothesis of "no difference" in feedback and therefore conclude that textual feedback reported significantly more usability defects than CTA, as the difference is positive $t(208) = 4.791$, $p=0.01$.

Author Keywords

Usability testing; Malaysian users;

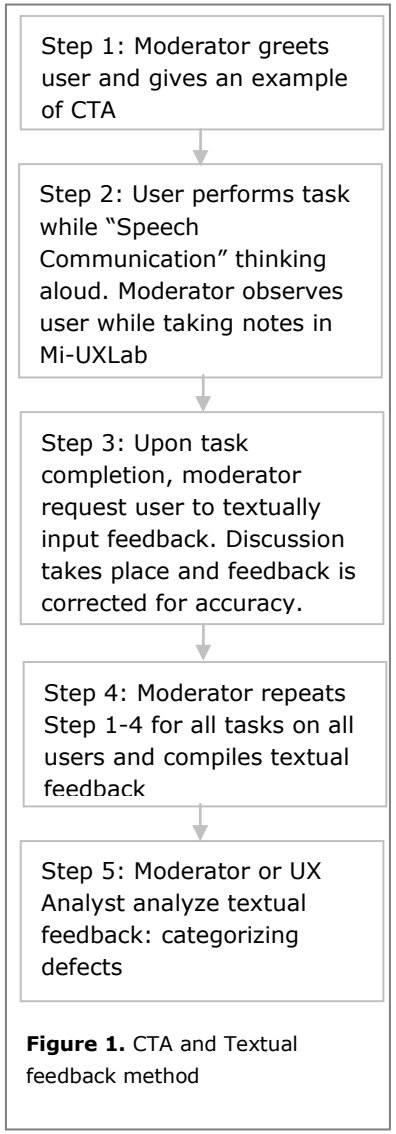

Figure 1. CTA and Textual feedback method

ACM Classification Keywords

Human-centered computing: Human computer interaction (HCI): HCI design and evaluation methods: Usability testing

Introduction

Think aloud protocol has been widely used in testing the usability of websites. There are various forms of think aloud protocols such as concurrent think aloud (CTA), retrospective think-aloud (RTA) and retrospective think-aloud with eye tracking (RTE). There are various advantages and limitation of each of these methods. Culturally, the moderators from MIMOS Usability/UX Lab found that generally Malaysian users are afraid that a failure of completion of a given task would reflect poorly on their performance rather than on the system or website being evaluated. This gives rise to users feeling reluctant to think-aloud during usability testing. As a result, insufficient usability defects are found when the CTA is used with Malaysian users. The aim of this study is to propose a new method known as textual feedback method for website usability testing and compare against CTA, particularly in terms of finding usability defects.

Related Work

Although [1] reported that RTE found three times more defects than textual feedback method, there are some practical issues with the eye tracker such as cost of ownership and stability of eye tracking software. The usefulness of CTA, which is one of one the dominant approaches in usability testing, was also debated by [2] due to its complexity and handling silences as compared when used with an eye tracker. However, study [3] found that performing usability testing with eye tracker has the highest cost of ownership as compared to other alternatives. Moreover, a regression analysis [4] shows that about 40 users are required from the eye tracker to obtain useful and interpretable results, which may not be feasible by all stakeholders. The secondary objective is to propose a practical and cost effective method for usability testing.

Concurrent Think Aloud Method

In this study, the CTA employed is similar to that described in [5, 6], also known as "Speech Communication"; at the start, the moderator greets the user and provides an example of how CTA is done. Once testing starts, minor verbal feedback by the moderator in the form of "um-hum" to keep the user talking is allowed; the moderator is also allowed to converse using a tone of questioning, or even reuse the last word verbalized by the user after 15 seconds of silence. This is shown as Step 1 and 2 in Figure 1.Based on speech-communication theory, it is important for the moderator to practice active listening and to be engaged in the conversation [7]. In contrary, study [6] has found that "Coaching" think aloud improves performance. As the websites evaluated are part of usability studies performed to measure efficiency, productivity and learnability, "coaching" think aloud is not suitable in our study.

Textual Feedback Method

In an e-commerce case study involving 6 subjects, textual feedback or also known as feedback capture after task (FCAT) were used to gather user's input [1]. This is achieved by prompting the user after the completion of each task to provide feedback on their experiences as shown in Step 3 in Figure 1. FCAT does not involve playback of any videos but instead rely purely on user's short-term memory of the experience.

There is a risk that users forget some of the issues they have faced. In this case, the role of the moderator is to remind or probe them of what they have mentioned when performing the tasks in Step 2. Since the moderator also has access to Mi-UXLab while hearing the CTA feedback during the task, the moderator would take notes during the various points of interest and later refer to the notes while encouraging the user to type in the feedback in their own words. One advantage of this method is that users will remember the issues that impacted them the most, hence it is expected that most of the important issues or words will be reported. This reduces false negative defects being reported.

Usability Defect Classification

The various websites evaluated in this study are intended to be used by consumers. Hence, an innovative and systematic usability framework comprising of objective performance and subjective image or impression were used as the first defect classification scheme [8]. This is known as the Han's classification scheme. The performance dimension is further classified into perception/cognition, memorization/learning and control/action. Meanwhile, the image/impression dimension consists of basic sense, image description and user's evaluative feeling. Those defects that do not fall in this scheme are classified as 'Others'. In addition to Han's, another scheme that is considered in this study involves evaluating the long term user experience. The Kujala's method of classifying long term system UX comprise of general UX, attractiveness, ease of use, utility, degree of usage over time [9]. Those defects that do not fall in this scheme are classified as 'Others'.

Methodology

We conducted a study involving 40 Malaysian users. Since each user conducted more than one task, a total of 209 usability testing tasks were obtained respectively using textual feedback method (condition 1) and CTA method (condition 2). These data were gathered and analyzed using SPSS vs 22. The analysis from both conditions comprise of verbalizations translated from CTA recordings, textual feedback converted to verbalization from Mi-UXLab and silences. Following Clemmensen et al. [11], we used Han's [8] and Kujala's [9] defect classification. The numbers of defects in the two methods were compared. This is to test the null hypothesis that there are no differences in defect classification between the two methods. During the study, the moderators and the users were not aware that the intention of this study is to compare the differences between the defects from the two methods. Thus there is no reason to suspect that the results are biased by the participant's expectations. Refer to Figure 1. From the above descriptions, the null and alternative hypotheses are as follows:

Ho: $\mu_T = \mu_{CTA}$ (There are no significant difference in the mean defects for textual feedback (μ_T) and mean defects from CTA (μ_{CTA}))

Ha: $\mu_T \neq \mu_{CTA}$ (There are significant difference in the mean defects for textual feedback (μ_T) and mean defects from CTA (μ_{CTA}))

Results and Discussion

From the 209 data set gathered, the test of normality violated the Kolmorov-Smirnov ($p < 0.01$) test. Anyhow, a violation of this assumption is of little concern to proceed with the paired-sample t-test, since

the sample size is large (more than 30) [10]. A paired-samples t-test was conducted to compare the number of defect classification for 209 usability tasks carried out using SPSS vs22 for textual method (condition 1) and CTA method(condition 2). Results reveal there were statistically significant difference in the defects between CTA method (Mean=2.68, SD=2.836) and textual method (Mean=4.09, SD=4.414); t (208) = 4.791; p = 0.01 (Table 2). Malaysian male and female also reported significant difference in the defects detected between the both methods; t (112) = 2.951; p=0.004 for male and t (97) = 3.939; p=0.01 for female.

Paired Sample Statistics	**Mean**	**N**	**Standard Deviation**
Textual	4.09	209	4.414
CTA	2.68	209	2.836
Paired Differences	1.402		4.230
Textual Male	3.10	113	2.130
CTA Male	2.27	113	2.836
Paired Differences Male	0.823		2.965
Textual Female	5.20	98	5.847
CTA Female	3.12	98	2.767
Paired Differences Female	2.082		5.232
Paired Differences	**t**	**df**	**Sig (2-tailed)**
Textual - CTA	4.791	208	< 0.01
Male	2.951	112	0.004
Female	3.939	97	< 0.01

Table 2. Paired sample statistics [10]

As hypothesized, we can reject the null hypothesis (Ho) of "no difference" in feedback and therefore conclude that textual feedback produces significantly more feedback than CTA, as the difference is positive. In other words, the textual method provides more meaningful feedback that could be translated to defects (mean 4.09) compared to the CTA (mean 2.68). In addition to this, a total of 854 defects were successfully classified using the textual method as compared to 561 for the CTA. This could be due to during the classification process, it was found that much of the verbalization during CTA involves the user reading aloud the task or trying to re-confirm with the moderator whether they are on the right track in performing the task or no feedback (silences). These feedbacks are classified as 'Others' and could not be classified under the Han's or Kujala's defect scheme. The maximum defect classified by a user is 32 for textual as compared to 14 for CTA. The minimum number of verbalization was zero which implies that users were silent during the testing. The frequency of silences recorded for CTA was more (21), as compared to 4 occurrences for the textual method.

Han's Defect Classification

Figure 2 and 3 shows the defect classification for textual feedback and CTA method. For the textual feedback, 61% of the defects are classified under performance, and 35% classified under image/impression. Only a small amount is classified under 'Other' category. The majority of defects for CTA was in the performance category. About one third of the defects are classified under 'Others' (32%). This was due to user reading aloud the task or trying to re-confirm with the moderator whether they are one the right track in performing the task or no feedback (silences).

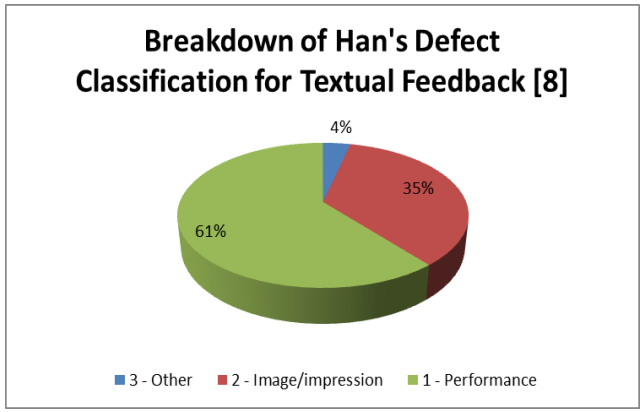

Figure 2: Han's defect classification for textual feedback

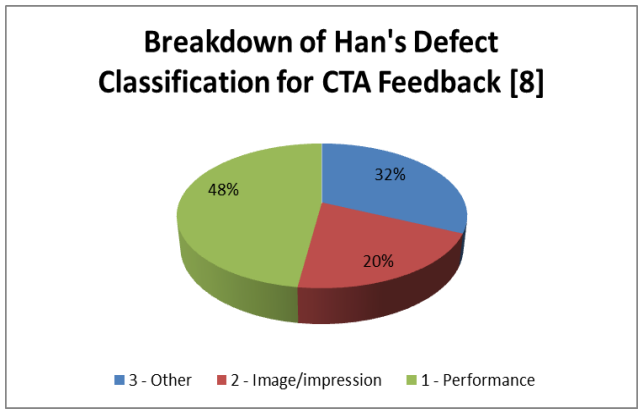

Figure 3: Han's defect classification for CTA feedback

Kujala's Defect Classification

For the textual feedback, the majority of defects are grouped under utility of the system (31%), followed by general relationship and user experience with the system (26%), ease of use (21%) and attractiveness

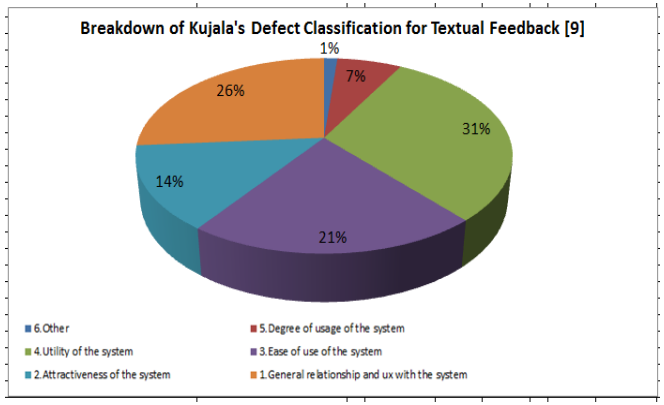

Figure 4: Kujala's defect classification for textual feedback

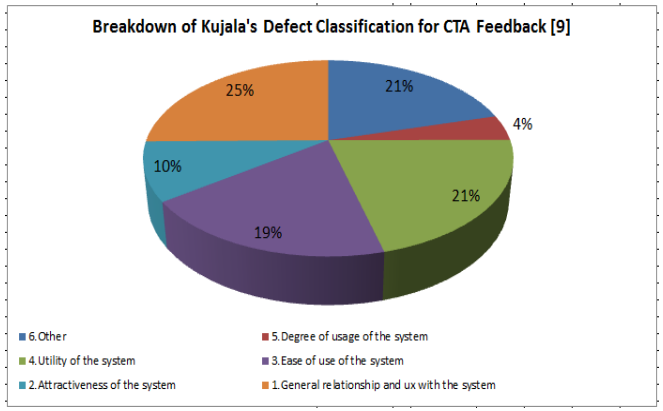

Figure 5: Kujala's defect classification for textual feedback

(14%). A small number of defects (1%) are classified as 'Other'. In contrary, for the CTA, almost one fifth of the defects (21%) were classified 'Other'. The reasons were similar to that in Han's.

Conclusion

This single blinded study revealed that there are significant differences between the textual feedback and CTA method. As hypothesized, the textual feedback method seems more culturally tolerant as it has the capability to detect more usability defects as compared to the limitations of the CTA for Malaysian male and female. This implies that the textual feedback method that is supported by the Mi-UXLab is more suitable to be employed for Malaysian users as compared to CTA. Some of the limitations are that there is quite a lot of variation between tasks and it would be beneficial to take this into account. It would also be of interest to take subject-specific explanatory variables, such as power status, into account. Future work include finding an appropriate model to address these limitations

Acknowledgements

We thank all the moderators from MIMOS Usability/UX Lab who performed the data gathering and offered the usage of Mi-UXLab 1.0 for the systematic data gathering.

References

[1] Goh, K.N., Chen, Y.Y., Lai, F.W., Daud, S.C., Sivaji, A., Soo, S.T. (2013). A Comparison of Usability Testing Methods for an E-Commerce Website: A Case Study on a Malaysia Online Gift Shop. Information Technology: New Generations (ITNG), 2013 Tenth International Conference , vol., no., pp.143,150, 15–17 April 2013

[2] Elling, S., Lentz, L., & de Jong, M. (2012). Combining concurrent think-aloud protocols and eye-tracking observations: An analysis of verbalizations and silences. Professional Communication, IEEE Transactions on, 55(3), 206-220.

[3] Sivaji, A, Tzuaan, Soo Shi; "Website user experience (UX) testing tool development using Open Source Software (OSS)," Network of Ergonomics Societies Conf. , 2012, IEEEXplore, pp.1-6.

[4] Nielsen, J. & Pernice, K. (2009). Eyetracking Methodology: How to Conduct and Evaluate Usability Studies Using Eye tracking. http://www.useit.com/eyetracking/methodology/eyetracking-methodology.pdf. Accessed June 15, 2011

[5] Boren, T. and Ramey, J. Thinking aloud: Reconciling theory and practice. IEEE Transactions on Professional Communication 43, 3 (2000) , 261-278

[6] Olmsted-Hawala, E. L., Murphy, E. D., Hawala, S., & Ashenfelter, K. T. (2010, April). Think-aloud protocols: a comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 2381-2390). ACM.

[7] Ward, N and Tsukahara, W. Prosodic features which cue back-channel feedback in English and Japanese. Journal of Pragmatics 32, (2000), 1177-1207

[8] Han, S. H., Hwan Yun, M., Kim, K. J., & Kwahk, J. (2000). Evaluation of product usability: development and validation of usability dimensions and design elements based on empirical models. International Journal of Industrial Ergonomics, 26(4), 477-488.

[9] Kujala, S., Roto, V., Väänänen-Vainio-Mattila, K., Karapanos, E., & Sinnelä, A. (2011). UX Curve: A method for evaluating long-term user experience. Interacting with Computers, 23(5), 473-483.

[10] Sheridan J Coakes, SPSS Version 20.0 for Windows, Analysis without Anguish, Wiley, 2013

[11] Clemmensen, T., Hertzum, M., Yang, J., & Chen, Y. (2013). Do Usability Professionals Think about User Experience in the Same Way as Users and Developers Do? Human-Computer Interaction–INTERACT 2013 (pp. 461-478): Springer.