ATLANTIS PRESS

Advances in Intelligent Systems Research (AISR), volume 149

International Conference on Science and Technology (ICOSAT 2017)

Usability Performance Evaluation of Information System with Concurrent Think-Aloud Method as User Acceptance Testing: A Literature Review

Yuditha Ichsani

Department of Informatics Engineering, Faculty of Science and Technology
Syarif Hidayatullah State Islamic University Jakarta
Email: yuditha.ichsani@uinjkt.ac.id

Abstract

In order to evaluate the user interface of information system in websites, softwares, applications, etc, some methods are needed to capture fully what is perceived by the user when accessing an interface, whether websites, mobile gadgets, desktop computers, and so on. One method that is growing in evaluating the usability of the user interface is a Concurrent Think-Aloud method. This method is an assessment method by including the verbal or speech opinion of the respondents and the respondents assessment results recorded during the evaluation process in real time. Since this method is conducted by end-user, then the result will have beneficial added value for information system improvement as good as already known method, Black Box Testing as one of User Acceptance Testing (UAT). One of weaknesses that can be found in this method is that some respondents may not show any expression or say to the evaluated website. This can be due to the character of the respondents themselves who tend to be less expressive. The advantages of this method are the developers of the information system can immediately understand what is experienced by end users rather than just using a questionnaire whose data collection tends to be less real time and make users tend to forget their real experience, and think-aloud method also can maximize end user's contribution to the UAT process.

Keywords: Usability, Concurrent think-aloud, User interface, User acceptance testing, Black box testing

1. Introduction

The relationship between humans and technology today is defined by human involvement with technology-based tools such as computers, mobile gadgets, and so on, known as human and computer interactions. The interaction is connected to a layer on the gadget that is the user interface. All interfaces that are directly used, need to be evaluated for the convenience of the user, either now or in the future.

What comprises good design? To be truly effective, good screen design requires an understanding of many things. Included are the characteristics of people: how we see, understand, and think. It also includes how information must be visually presented to enhance human acceptance and comprehension, and how eye and hand movements must flow to minimize the potential for fatigue and injury. Good design must also consider the capabilities and limitations of the hardware and software of the human-computer interface. Excess learning requirements can also become a barrier to users achieving and maintaining high performance and can ultimately influence user acceptance of the system. [1].

The International Standards Organization (ISO 9241-11) defines usability as the extent to which a product can be used by a particular user to obtain a particular goal with effectiveness, efficiency, and satisfaction in the context of use. These three factors can not be separated in measuring performance usability. The Usability Professionals Association (UPA) provides usability definitions that focus more on the product development process. Usability is an approach to product development that combines user feedback through the development cycle to reduce costs and create products and tools that meet user needs [2]. The next definition is put forward by References [3], that Usability means ensuring that something works well for a particular purpose without the user becoming discouraged.

In conducting user interface evaluation, we need a method that can fully capture what user feels when accessing a website interface, mobile gadgets, desktop computers, and so on. One method that is evolving in evaluating user interface usability is the Concurrent Think-Aloud (CTA) method. This method is an assessment method which includes the verbal opinion of the respondent and the result of respondent's assessment is recorded during a real time evaluation to capture their experience.

The term of "think-aloud protocol" refers to a type of research data used in empirical translation process research. The data elicitation method is known as "thinking aloud" or "concurrent verbalization", which means that subjects are asked to perform a task and to verbalize whatever crosses their mind during the task performance. The written transcripts of the verbalizations are called think-aloud protocols (TAPs) [4].

The purpose of this paper is to propose Concurrent think-aloud method as information system usability performance evaluation to be implemented in User Acceptance Test (UAT) process. The research method is

literature study and also comparison between the researcher's experience and relevant literatures, so there is a strong reason to achieve the above objective, which in turn can maximize the collection of respondents preference data as end users of an information system.

2. Research Methods

A. Observation

This paper also revisit researcher's previous experience in conducting think-aloud method. In 2012 and 2014, the author have implemented the Concurrent Think-Aloud method as one of the methods of evaluating website usability. Provincial government websites were the research objects in 2012 research and State University Websites were the research objects in 2014 research. There were 31 respondents in 2012 research which consist of 23 females and 8 males. In 2014 research, there were 26 respondents which consist of 8 females and 18 males. They were chosen since they are all active internet users (internet literated). At those research period, observations were made by observing changes in gesture and respondent expression while performing some scenarios provided by the researcher. The observation was done carefully but it is endeavored that the respondent will not be disturbed during the observation process, such as by taking the position beyond the reach of respondent's view. In addition, gestures and respondent's expressions in both researches were also recorded using cameras on laptops (web camera) and Camtasia version 7.1 in 2012 research, version 8 in 2014 research as screen capture software. Fig. 1 shows Camtasia's display in analyzing captured video, which has laptop screen and respondent's front face captured by web camera.

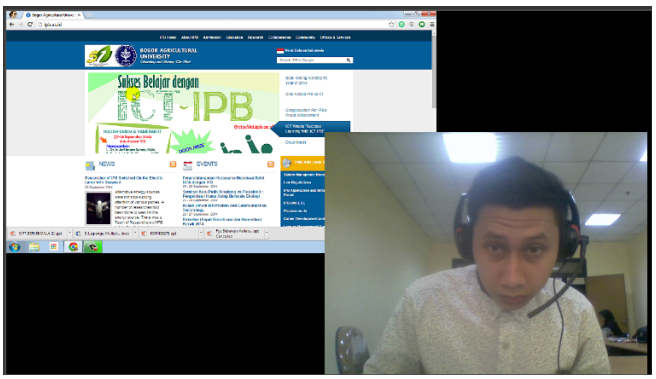

Figure 1. Camtasia's display as screen capture software

CTA was chosen by researcher is because the researcher will also calculate the amount of time and number of steps taken by the respondent in completing the given scenario. This is in line with [5] which states that one important reason to avoid asking participants is to think-aloud is when you are measuring time on tasks. Thinking aloud slows performance significantly. Some practitioners also avoid retrospective review because it gives some participants opportunity to revise and rationalize their behavior rather than simply reporting on what happened and why. In addition, the other reason of researcher chose this method is because researcher wanted the respondents will be more free to express their opinions without any interference from the researcher.

B. Literature Study

The researcher conducted literature study to find background theory that support the proposed idea and related researches from some sources. Research reports, thesis, dissertation, reference textbooks, e-books, journal articles and relevant websites were used as the paper literature sources.

C. Related Research

In 2002, Dorman and Markopoulos conducted think-aloud method to childern aged 8-14. The results of [6] study show that the think-aloud protocol helps identify most usability problems and suggests that girls thinking out loud report more usability problems than boys.

Furthermore, think-aloud or think out loud method had been conducted by undergraduate students. The students finally learn how to perform formative evaluation of systems, and the researchers believe that they are more likely to design usable and useful systems in the future. More importantly, one goal the instructor set for these students has been accomplished; they successfully learned to conduct a usability study in which they systematically contributed to the on going development of an software tool through usability engineering [7].

Other research about concurrent think-aloud describes an experiment that compares concurrent and retro-spective think-aloud protocols for a usability test of an online library catalogue. There were three points of comparison: usability problems detected, overall task performance, and participant experiences. Results show that concurrent and retrospective think-aloud protocols reveal comparable sets of usability problems, but that these problems come to light in different ways. In retrospective think-aloud protocols, more problems were detected by means of verbalization, while in concurrent think-aloud protocols, more problems were detected by means of observation. Moreover, in the concurrent think-aloud protocols, the requirement to think-aloud while working had a negative effect on the task performance. This raises questions about the reactivity of concurrent think-aloud protocols, especially in the case of high task complexity [8]. Furthermore, it can be concluded that concurrent think-aloud method will be more useful for evaluating tasks in moderate or low complexity, such as finding information and create an account., as stated by [8], that directions offered for think-aloud research often state that the researcher should formulate tasks with a moderate difficulty, so that participants are not inclined to follow an automated working process, but will also not be burdened with a cognitive load that is too high.

The limitations of the think-aloud method of data collection are thoroughly explored in this paper. It is essential for one who desires to use this method of data collection to be fully informed of these limitations in order to elicit useful, relevant and sufficient data. However, [9] firmly believe that what is added to the research data through the use of this research method far outweighs these stated limitations. [9] concur that while it is not

claimed that think-aloud data provides a complete insight into the human mind, it certainly is a useful tool available to the researcher.

In [10] research, the cultural impact on thinking aloud usability testing is investigated. Evaluators and users with the same and different cultural backgrounds were invited to attend the study in order to examine the extent to which their cultural backgrounds impact the thinking aloud usability testing. The evaluators' and users' cultural backgrounds may influence both usability problems and communications. In this research, Denmark is selected to represent the Western culture, and China is selected to represent the East Asian culture. Accordingly, the usability tests were conducted in Denmark and China with Danish and Chinese evaluators and users, respectively, in order to investigate the research question. This research found that communication seems important for usability tests in both Western and East Asian cultures. Western users give more negative comments than do East Asian users. Users, especially East Asian users, give more culture related comments to foreign evaluators than they do to local evaluators. In order to find usability problems, usability practitioners may need to pay attention to users' negative comments, suggestions and questions. In order to encourage the users to talk, the evaluators need to give lots of affirmative responses. In order to understand the users' problems, the evaluators need to give digging deeper probes. For the potential problems that the users do not notice, evaluators may need to direct the users' attention on those issues and to get feedback from the users.

One difference that added by this paper from other related research is the facial expression and body gesture included in think-aloud protocol, not only verbal protocol. That is because the observation conducted was also captured by web camera to become complete audio-video file not only by researcher in order to make respondents feel comfortable to do whatever they think must do in completing the tasks. This also in line with [11] that the current discussion concerning the think-aloud method and general litterateur on the method has as far as can tell a focus point on the test setup (i.e. use of video etc.).

3. Discussion

A. Usability Evaluation

Respondents were placed in an observation area equipped with portable computers connected to the Internet using a modem and installed screen capture software, along with web cameras, headsets and a set of questionnaires. Web cameras mounted on portable computers are used to capture facial reactions and activities performed by respondents. Respondents were asked to use a headset and were asked to say the action they were doing, the thought of it, the feelings experienced during the continuous evaluation to be recorded by the microphone on the headset, then the resulting transcripts were analyzed for dominant positive or negative reactions. This method is called Think Out Loud [12]. The other terminology used in this paper is Think-Aloud.

Similar with [12], Barnum in [13] also stated that the usability test goal is to improve the usability of a product. The participants represent real users and they do real tasks. The researchers observe actions and record what the participants say and analyze the findings, diagnose problems, and recommend changes.

The researcher also gave questionnaire to be filled by respondents as written data sources and it also said "When conducting a research scenario, please express it in spoken language (like saying "Found it!" when the information is found)" on the beginning of questionnaire's introduction. Participants may be asked to give their comments either while performing each task ('think-aloud') or after finishing all tasks (retrospectively). When using the 'think-aloud' method, participants report on incidents as soon as they happen. When using the retrospective approach, participants perform all tasks uninterrupted, and then watch their session video and report any observations (critical incidents) [14]. Figure 2 and 3 shows the observation environtment while doing usability evaluation research in 2012 and 2014 respectively.

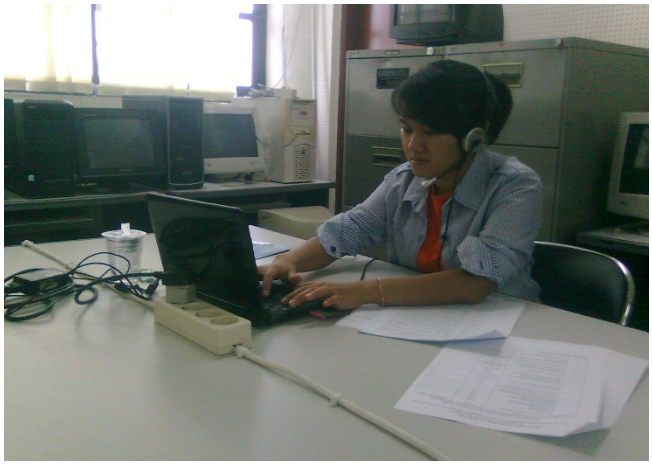

Figure 2. Websites usability evaluation with concurrent think-aloud method in 2012 research

Figure 3. Websites usability evaluation with concurrent think-aloud method in 2014 research

Table 1 shows some examples of 26 respondents speech transcript in Bahasa Indonesia (Think-Aloud Protocol), gestures, and facial expressions using Think-Aloud Method in 2014 research.

Table 1. Example of respondents feedback as think-aloud protocol in 2014 research

No	University Abbreviation	Positive Comments and Expressions	Negative Comments and Expressions	Conclusion
1.	*ITB*	*1. "Wow"* *2. "Wow keren"* *3. "Wah keren"* *4. "Keren"* *5. "Bagus"* *6. "Baik banget"* *7. "Keren ITB"* *8. "OK"*	*1.* (Surprised) *2.* (Confused)	Positive Feedback is more dominant
	Total	***8***	***2***	
2.	*IPB*	*1. "Wow, baik sekali"* *2. "Lumayan IPB"* *3.* (Hummed)	*1. "Ah ribet deh"* *2. "Kok jadi lama?"* *3. "Ah kok nggak ada?"*	Balance between positive and negative feedback
	Total	*3*	*3*	
3.	*UGM*	*1. "Wow"* *2. "Bagus"* *3. "OK"* *4. "Keren"*	*1. "Yah cuma ada di Home"* *2. "Ada tapi pake search"* *3. "Yah mana nih?" (Confused)*	Positive Feedback is more dominant
	Total	*4*	*3*	
4.	*UI*	*1. "OK"* *2.* (Hummed) *3. "Wow"*	*1. "Kecil banget tulisannya"*	Positive Feedback is more dominant
	Total	*3*	*1*	
5.	*UNDIP*	-	1. (Laughed and confused) 2. (Sighed) *3. "Kiri kanan bingung"* 4. *"Lah?"* (Surprised) 5. (Grimmed) *6. "Banyak banget dikliknya"* *7. "Yah, kebuka lagi"* *8. "Ya Allah!"* *9.* (Sighed)	Negative Feedback is more dominant
	Total	*0*	*9*	
6.	*UNHAS*	*1. "Bagus juga UNHAS"* *2. "Bagus tidak perlu pakai Search"* *3. "Bagus juga"* 4. (Smiled)	*1. "Yah, lagi maintenance"* *2.* (Surprised)	Positive Expression is more dominant
	Total	*4*	*2*	
7.	*UIN Jakarta*	*1.* (Hummed) *2.* (Hummed 2) *3.* (Whistled)	*1. Iih?!* (Surprised) *2. Beda lagi* *3. Lain ini mah!* 4. (Long Sighed) *5. "Wah dimana kontak?"* *6. "Dimana ? Kok nggak ada?"*	Negative Feedback is more dominant
	Total	*3*	*6*	

Notes:
" " : responds in comment or verbal expression
( ) : responds in body gesture and facial expression

Based on Table 1, it can be seen that responses obtained for all websites in the form of words, face expressions, or gestures are quite small when compared with the number of respondents (26 people). Therefore, the recapitulation result of the reaction resulted from some respondents only, because most of the respondents did not show positive or negative feedback when accessing and evaluating the website.

The determination of feedback types that has a positive or negative tendency is based on a general principle that is clearly shown by a person when dealing with a thing. The researcher categorize positive feedback such as, compliment, satisfaction statement, amazed, and comfort condition expression. The 'humming' and 'whistling' condition were categorized as positive feedback since it represent comfort condition and this is in line with [15] who said that behavioral ratings revealed that happy music without lyrics induced stronger positive emotions than happy music with lyrics. Meanwhile, the negative feedback included in this paper are criticism, discomfort statement, confusion, and unexpected surprise. If the feedback in the form of words shown is neutral and the facial expression is also neutral (known as poker face), then it is not used. But if the spoken word is neutral but facial expression is not neutral, then the response is categorized to be positive or negative.

Based on some observations conducted by the researcher, some respondents hesitate in showing their expression or feedback, although it has been convinced before by researchers to feel free to show their feedback. This is thought to be due to the diversity of the nature of the respondents, whether including introverts or extroverts, which requires study in the field of psychology.

The response or reaction given by the respondent during the usability test with the concurrent think-aloud method is more in terms of words or phrases consisting of two to five words rather than in the form of a facial expression clearly indicating joy (like) or dislike.

B. Advantages and Disadvantages of Concurrent Think-Aloud Method

There are two advantages of think-aloud method based on [1], i.e. utilizes actual representative tasks to be conducted by end-users and provides insights into the

user's reasoning in term of system evaluation. Valuable insights into why the user does things are obtained in order to get complete user's point of view.

From the researcher prespective, the advantages of this method are the developers of the information system can immediately understand what is experienced by end users rather than just using a questionnaire whose data collection tends to be less real time and make users tend to forget their real experience. If the respondent fill in the UAT questionnaire after the evaluation conducted, there is possibility of forgetting the previous experience rather than say the experience itself while having the evaluation. Furthermore, the problem finding process can be accelerated by having manual blank timeline in ruler form and give the sign or stamp for every comment made by respondent, so the researcher can analyze the video directly to a certain parts.

The second advantage is concurrent think-aloud method also can maximize user's contribution to the UAT process, such as collecting expressions, verbal and written comments at a time. Finally, this method is expected to improve satisfaction of information system end users. This statement also supported by [16] who stated that think aloud protocols are becoming more common in educational research due to the richness of data that potentially can be derived from the methodology.

It is in line with [11] who thought that there is a good possibility that the method is used in almost all aspect of testing the usability of applications and interfaces. From investigating the mental model of the user through "heavy" propping to the assessment of usability as well as usability performance data in the form of efficiency and effectiveness.

The principle weakness that can be found in concurrent think-aloud method from researcher's experience is that some respondents may not show any feedback or comment to the evaluated website. The screen capture software (such as Camtasia) can provide audio track that can be analyzed after the test conducted, and the researcher can only analyzed the active audio area as the difference is seen in Figure 4 and 5.

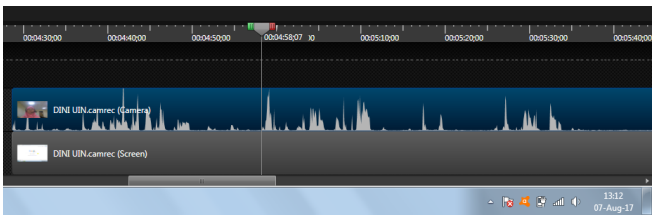

Figure 4. Active respondent' audio track

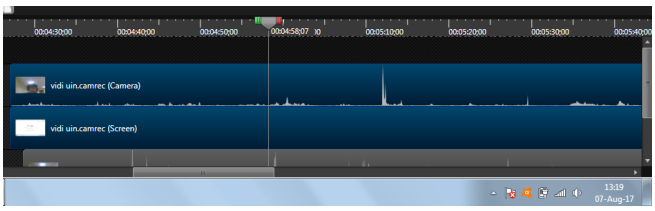

Figure 5. Passive respondent' audio track

Furthermore, [1] also stated that it may be difficult to get all people to think out loud. This problem can happened due to the character of the respondents themselves who tend to be less expressive. The first solution proposed are to have balance amount of female-male respondents in order to have a better and more active concurrent think-aloud respondents that is in line with [6], and also respondent's background, s which is in line with [10]. The second solution proposed in this paper is to have an interview with all participants before usability evaluation conducted in order to find out the personal characteristic of them and choose only those who are classified as active in speaking and expressive in issuing their personal opinions. [10] also found in his research that evaluators' and users' cultural backgrounds do impact the usability testing, but the influence is not the same for participants with different cultural backgrounds. The research implies that users' different cognitive styles and communication orientations may have different impacts on the thinking aloud usability testing result and the usability testing process. The third solution proposed in this paper is to have the written sign that says "Please keep talking about what you feel" at a place that is on the respondent's visibility, for example on top border of laptop screen in order to keep reminding the respondents for talking about what their thoughts. This is also in line with [16] who stated that neutral cues such as "keep talking" encourage subjects to think-aloud but do not bias the data by adding external ideas to the internal processes of subjects.

The fourth solution stated by [11], that is the different persons (researchers) involved in the test should behave in order to get the best data possible. This could for example be how to get the test person to keep speaking aloud and how the person "interviewing" the test person should behave, in order not to influence the way the test person would solve a problem or perform at task. It is also can be said that the researcher can give interview to repondents during the test to to encourage the respondents in expressing their thoughts and feeling through comments and gestures. Regarding this kind of test, the researcher tend to have the observation outside respondent's sight since the respondent might feels to have an obligation in making good comments, as mentioned in [14] that participants tend not to voice negative reports.

The second weakness found by researcher is the process of collecting data from the respondents can make them feel bored and it will impact to their mood in doing the evaluation. As a comparison, in 2012 research, 31 respondents must conduct usability evaluation for 3 provincial government websites each, but in 2014 research, 26 respondents must conduct usability evaluation for 7 state university websites each. Therefore, this can be solved by limiting the scenarios and trying to get a short duration of the video by only conduct the test for some of the most crucial tasks after they were found from previous test by system developer, that is White-Box testing. This also in line with [1] who gave a guideline to conduct think-aloud method such as the researcher can develop core or representative task scenarios, or scenarios of proposed tasks of particular concern and limit a session to 60 to 90 minutes per person. The next proposed solutionto make think-aloud evaluation more efficient to encourage

respondents in expressing what they feel, not what they do in order to achieve their satisfaction.

4. Conclusion

The concurrent think-aloud method has a good opportunity to become UAT support to gain more information from respondents in certain tasks, such as moderate and low complexity tasks, such as finding information and create an account. Some advantages of this method are the developers of the information system can immediately understand what is experienced by end users rather than just using a questionnaire whose data collection tends to be less real time and make users tend to forget their real experience, and concurrent think-aloud method also can maximize user's contribution to the UAT process.

The principle weakness that can be found in this method is that some respondents may not show any feedback or comment to the evaluated website. This can be due to the character of the respondents themselves who tend to be less expressive. The solution proposed are to have balance amount of respondents gender and backgound, to have an interview with the participants who willing to become usability evaluation respondents, and the researchers must be active enough to encourage the respondents in expressing their thoughts and feeling through some comments and gestures. The future research can provide an improved concurrent think-aloud method to compare usability evaluation in websites, dekstop and mobile phone applications, then redo the evaluation after those system improved based on the previous test. Furthermore, the research scope can be expanded by using eye-tracking software and speect-to-text software in Bahasa Indonesia.

Acknowledgement

This work was supported in part by Center of Research and Publishing (Puslitpen) Syarif Hidayatullah State Islamic University Jakarta in 2014 Research Grant.

References

[1] Galitz WO. 2002. An Introduction to GUI Design Principles and Techniques. New York: John Wiley and Sons.

[2] Tullis T, Albert B. 2008. Measuring User Experience Collecting, Analyzing, and Presenting Usability Metrics Interactive Technologies. Massachusetts: Morgan Kauffman.

[3] Krug S. 2006. Don't Make Me Think: A Common Sense Approach to Web Usability 2nd ed. California: New Riders Press.

[4] Jääskeläinen R. 2010 Think-Aloud Protocol in Handbook of Translation Studies. Edited by Yves Gambier and Luc Van Doorslaer. Page 371. Amsterdam: John Benjamins Publishing Company

[5] Rubin J, Chisnell R. 2008 Handbook of Usability Testing, Second Edition: How to Plan, Design, and Conduct Effective Tests. Indianapolis: Wiley Publishing, Inc.

[6] Donker A, Markopoulos P. 2002. A Comparison of Think-Aloud, Questionnaires And Interviews for Testing Usability with Children. Proceeding of HCI 2002. 305-316.

[7] Czajkowski MF, Foster CV, Hewett TT, Casacio JA, Regli WC, Sperber HA. 2001. Student Project in Software Evaluation. ITiCSE 2001 6/01 Canterbury, UK.

[8] Van den Haak MJ, De Jong MDT, Jan Schellens P. 2003. Retrospective vs. Concurrent Think-Aloud Protocols: Testing the Usability of an Online Library Catalogue. BEHAVIOUR & INFORMATION TECHNOLOGY, SEPTEMBER–OCTOBER 2003, VOL. 22, NO. 5, 339–351

[9] Young KA. 2005. Direct from the source: the value of 'think-aloud' data in understanding learning. Journal of Educational Enquiry, Vol. 6, No.1, 2005, p19-33

[10] Shi Q. 2010 An Empirical Study of Thinking Aloud Usability Testing from a Cultural Perspective. DISSERTASION. LIMAC PhD School, Programme in Informatics, Department of Informatics. Copenhagen Business School (CBS)

[11] Ginger-Mortensen R. 2007. Do words get in the way of (better) usability? A theoretical-conceptual analysis of the think-aloud method in relation to verbal overshadowing. THESIS. Department of Informatics. Copenhagen Business School (CBS)

[12] Mayhew DJ. 1999. The Usability Engineering Life Cycle. San Francisco: Morgan Kauffman.

[13] Race C .2016. "38. Usability Testing". Sexy Technical Communications. 38. http://digitalcommons.kennesaw.edu/oertechcomm/38

[14] [USDHHS] US Department of Health and Human Services. 2004. Research-Based Web Design & Usability Guidelines. www.au.af.mil/pace/handbooks/usability_guidelines.pdf.

[15] Brattico E, Alluri V, Bogert B, Jacobsen T, Vartiainen N, Nieminen S, Tervaniemi M. 2011. A Functional MRI Study of Happy and Sad Emotions in Music With and Without Lyrics. Frontiers in Phsycology. December 2011, Volume 2, Article 308.

[16] Johnstone CJ, Bottsford-Miller NA, Thompson SJ. 2006. Using the Think Aloud Method (Cognitive Labs) To Evaluate Test Design for Students with Disabilities and English Language Learners (Technical Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.