Introduction

What is Usability and Usability Testing?

The definition of usability can vary depending on the researcher. Usability is commonly defined as a measurement of effectiveness, efficiency and ease of use of an application, product or website in relation to the user interface. Jakob Neilsen will define usability as a quality attribute of five components: Learnability, efficiency, memorability, errors and satisfaction. (http://www.useit.com/alertbox/20030825.html). Overall, usability refers to how well people learn and use a product to achieve their goals and how satisfied they are with that process.

Relative to software or internet development, usability testing has been growing slowly. As interacting with computers becomes daily routine there has been more emphasis on creating highly usable products. Some products will live and die based on how easy they are to use and how satisfying they are to use. Making something usable can be a long and difficult process but well worth it in the end. Anything that a person can interact with can be usability tested. Many different methods and procedures have been formulated to analyze the user experience. These various testing methods are part of an iterative process that occurs in a products development cycle. The iterative process is part of user centered design methodology. It is important to note that usability testing is not showing the user a product and asking if they understand it, doing that could cause a different usage pattern of the product than would normally happen. The goal is to observe the user interacting with the product as naturally as possible to ensure that the interaction is the same as when alone. It is common that participants in usability studies will try to give the moderator[1] the answer they believe they are looking for, even if that is not what they really think and this will skew the results or recommendations.

What is User Centered Design?

User centered design, UCD, is a methodology for the process of developing and designing new products for the market and the best way for employing usability. The core idea behind UCD is developing for the user and incorporating the user in the development process. For example, when creating a new website, the incorporation of UCD methods involves finding members of the public that would be ideal users of the product and asking them for opinions and feedback at various stages of development. UCD is a method growing in popularity; major companies like IBM or Google focus on using some UCD methods to produce better products. UCD is a process of repeatedly testing a product while it is in development with specific users using a variety of usability testing methods. Usability testing is suitable for any stage in a products life but in general the earlier and the more often the better.

Brief Introduction to Usability Testing Methods

There are a variety of usability testing methods and procedures, in fact more than enough to fill this paper. The focus of this report however is think aloud protocol, which is normally incorporated as a part of other types testing methodology and will be explained later in the report. There are conflicting views of when to use the think aloud protocol and how effective it is on testing a product. A short list is provided here as an introduction to the various usability testing methods and how thinking aloud is incorporated if at all.

- ***Focus groups***
    - Focus groups involve gathering a group of approximately 8 to 12 people to discuss the ideas behind the product and gage user's reactions and attitudes towards it. It is conducted as a guided discussion where the moderator wants to make sure that they conversation is on task and the ideas are flowing. Focus groups are usually conducted early in product life cycle while looking at concepts or design ideas.
- ***Card sorting***
    - Card sorting uses various cards of information or concepts to understand how users think about the content and how they would organize it within the product. It is usually conducted in a one on one interview where the participant is asked to group concepts or information in a logical order to them. Card sorting can be used at any stage of the development cycle although it is regularly conducted early. Often when participating in a card sort the user is asked to think aloud about their decisions.
- ***Interviews (Contextual or Individual)***
    - The user and the moderator discuss the product, often using a set of predetermined questions. With the interview method, the moderator is able to get a deep understanding of the participant and is provided an opportunity to discuss interesting concept in depth. Contextual interviews take place where the participant would actually be using the product, often times at the participant's home. Interviews can be used at any stage of the product development depending on what you need to learn.
- ***Prototype testing***
    - Prototype testing involves getting the user to interact with a prototype of the product and observing their behaviors and frustrations. Prototype can be working model or mock-ups on paper where the participant would indicate what they wanted to do next to the moderator. This type of testing is done using the early versions of the product and is often helpful with navigational problems. Participants are often encouraged to think aloud during these studies primarily so if the moderator has to move to the next screen or part of the product manually, they are able to understand and move as seamlessly as possible.
- ***Lab studies***
    - Lab studies are conducted in a usability lab, when a participant visits the company and actively uses the moderator's equipment (usually computer) to test the working product. The moderator is able to see how the user will interact with the product in real time and where they will find faults. There are many variations to add to the lab study including eye tracking etc. Lab studies are usually conducted with a working model of the product and in

later stages of the development cycle. Think aloud protocols and commonly practiced during lab studies to better understand the user.

There are different variations of all these studies and there are many more not mentioned. Most companies will do a variety of testing based on the product and the stage of development. As stated before testing is an iterative process of the development of the product, once a test is complete changes are made and testing begins again.

Thinking Aloud Protocol

Previous Research on Think Aloud Protocol

Asking the participant to verbalize their thoughts or to think aloud has become common in usability studies, to the point that it is simply understood in some circumstances. There are, as with any protocol, times when it should be used and times when it should not. Any protocol or method such as think aloud, requires research and proper implementation before it can be considered common practice. There are a variety of views on how and when to use the think aloud protocol and varying opinions about the merits and trade offs of it. Think aloud protocol began in cognitive science and psychology. Most commonly sited as the theoretical base of the protocol is Ericsson and Simon for their work in 1984 on "Protocol Analysis: Verbal Reports as Data" [2]. Much of the book however focused on applications in cognitive science. Namely, the three forms that a participant's verbalizations can be formed in and which ones are suitable for study. Ericsson and Simon also outlined an elaborate methodology for conducting the think aloud protocol. Boren and Ramey [1] later researched the think aloud protocol in relation to usability studies and discovered that the methodology that Ericsson and Simon described was commonly referenced but in practice was rarely followed. They continued to analyze the way usability researchers implemented the think aloud protocol and developed a new methodology for the use of usability studies. The new methodology is a combination of what is used in practice and what Ericsson and Simon suggested. There were three main areas of change each increasingly opposing to Ericsson and Simon's methods. Upon reviewing these two methodologies Krahmer and Ummelen conducted a comparison study to determine effectiveness of each for usability testing [3]. By strictly conforming to the guidelines of each protocol, they were able to observe that the process of thinking aloud was not affected by either approach and task performance did not differ. However, the participant's following the Boren and Ramey approach were less lost and completed more tasks. Overall, Krahmer and Ummelen noted that the evaluations of product in question did not differ based on the protocol used.

Norgaard and Hornbaek take a critical look at the way think aloud is conducted and in particular how it is analyzed by usability professionals [6]. They note that there are realities of think aloud protocols that have large impacts on usability testing. There of course still exist areas where think aloud creates conflict during a usability study. Having a participant think aloud while working with any product is not natural behavior [9]. Therefore, think aloud protocol is not recommended for usability studies that intend to review certain metrics such as time on task[2]. Rosson and Carroll explain this as one of the tradeoffs of using a think aloud protocol [9]. One proposed idea of how to address the latency is to conduct the think aloud portion after the task is complete; this is referred to as retrospective think aloud (RTA). There are many studies about the effects of RTA verses concurrent think aloud (CTA) with varying results. A study in 2003 by van den Haak and de Jong indicated that the two methods produced similar numbers and types of problems. Notably, the CTA verbalizations noted more problem detections where the RTA verbalizations seemed to be more substantial [10].

Think Aloud in Practice

There are many papers about think aloud protocols in practice, how they are conducted and their relative effectiveness. Papers can include comparisons of think aloud verses silent, or concurrent verses retrospective, or even two different implementations of think aloud. There are various types of usability studies when having the participant think aloud can provide important insight into the product that may have been over looked before. Many times however the verbalizations are used in conjunction with video to make supporting arguments for the usability recommendation. Audio recordings are played back for the customer of the usability study (commonly an engineer, designer or product manager) so that they can gather a better understanding of what could be changed on the product. There are important times during the products life when usability testing with the participant's opinions are increasingly valuable. When a usability researcher is trying to test aspects of a product such as navigation and understanding having opinions matters. Either the participant can be allowed open-ended exploration or task based scenarios, in both cases the verbalizations are extremely helpful for the moderators understanding. If a product is in the early stages of development sometimes user's quotes can lead the product to different features that were not considered before, at this point it seems to help conceptually with the product. In later stages of the product cycle, it can help indicate areas of confusion for the user, where the vocabulary and mental model of the user and developer differ.

Sometimes usability professionals also want to find out how long it takes to complete a task using the product. In this case, it is not advisable to use a concurrent think aloud protocol because of the fear that it will skew the time on task of the participant. If the participant is being pushed to think aloud in a manner that is not natural to them, it is perceived that they will take longer than if they were working naturally. Because it is not always known when to use the think aloud protocol it is less often that an actual methodology is described regarding how to use it.

Methods

Best Practices

Every usability practitioner has different ways of conducting their usability study from instructions to mannerisms to expectations. It is rare that the outline for the think aloud protocol, either Ericsson and Simon or Boren and Ramey, is strictly followed; perhaps it is not crucial that they are carried out perfectly. Listed here is a compilation of some of the best practices for carrying out a concurrent think aloud usability lab study.

- ***Lab setup***
    - In any typical usability lab, there should be a device to record audio and screen capture if you are working with a computer system. A video of the participant’s facial reaction is sometimes interesting to record and can help further reinforce recommendations of the study afterwards. The participant is likely to feel less awkward about thinking aloud if the moderator is in the room with them. It is important for the moderator to try not to influence the participant by means of body language or tone of voice; this may be easiest to accomplish by sitting slightly behind the participant.
- ***Before the study takes place***
    - As with any usability study is it incredibly important for the participant to be as comfortable as possible and for them to understand any instructions they are given. The moderator should not talk too fast and always check to the participants understanding. The moderator should ask the participant to think aloud during the study and give examples of what they mean. At the same time, the participant should understand that all usability studies are a test of product and not the participant. Invite them to be open with their ideas, comments and thought processes during the study. It is important to inform them that while they should ask questions the moderator may not be able to answer them for various reasons. Some moderators will give examples of suitable questions that they will gladly answer, such as when a task is unclear.
- ***During the study***
    - In order to keep the participant comfortable and talking during the study the moderator should be aware of their surroundings. It is important for the moderator to stay neutral and keep all comments and body language to a minimum. Some find it best not to take elaborate notes while moderating because of the disruptive effects that it has on the participant’s train of thought. When the participant asks questions, the moderator must carefully consider how to answer them, most questions can be answered with a question that will encourage the participant to keep talking and work through the problem. An identified problem with the think aloud protocol is how often to interrupt and what to say when the participant falls silent. Boren and Ramey encourage the use of natural acknowledgement tokens such as “Mm Hmm” and “Oh yeah” that will subtly remind the participant to keep talking. It is however difficult to quantify how often the moderator should remind the participant, most will ask questions or remind them after following natural conversation cues.

- ***After the study is complete***
    - Once the participant has completed all of the tasks, ask if they have any final comments or questions. In some usability studies, it has been profitable to let the participant speak freely about the product they have just been using. In this way some participants do not feel constrained by the tasks they were working. A caution against this idea is that participants may be not speaking naturally about the product and could be looking to give the answers that they think the moderator is looking for. Once the participant has left the room, the moderator can check that the recordings worked and make short notes about the session[3] for later use. In addition, if there were any other people observing the study, this is the time to discuss with them and gather their notes about the session.

Think Aloud Analysis

There are different ways to analyze the think aloud data in theory and in practice. Looking back to the Ericsson and Simon methodology of think aloud they would advise transcribing each session with the participant and encoding it to allow gathering metrics without biases. In common practice of usability testing, this type of analysis in not practical and not revealing of issues and bugs of the product. There often is criticism about how analysis of any usability study should take place, as well as how to create recommendations for the product. What has been observed by some in practice regarding analysis may not work for all usability practitioners. Some intuition and subjective interpretation is often mixed into the analysis process. Much work still needs to be done on defining procedures for fast-paced analysis for usability studies in industry.

Some ideas to consider when analyzing think aloud sessions of a usability study include reviewing the audio or video and making notes that are more detailed. Be sure to listen for comments about the product as well as tone or inflection in the voice when talking about the product. Note the user’s reaction when they are lost or confused, where do they place blame? Can they explain what went wrong or do they know that something is wrong? While reviewing the audio or video it is helpful to record time stamps of important quotes or actions, as these will prove helpful later. From the notes of all sessions themes should arise where the moderator can then draw parallels from and create a set of findings. Concordantly, findings will lead to recommendations for the product.

Recommendations and Forward Thinking

Many usability practitioners have recognized the value of gathering quotes and opinions about a product while testing it. Most will call this the think aloud protocol, even if their method and implementation differ vastly from the documented protocol of Ericsson and Simon, which is most commonly referenced. Participants of the usability study are often given few instructions on what thinking aloud actually requires them to do and therefore can become confused. Think aloud can help just about any usability study but it seems to be best at navigational[4] and exploration[5] studies. In the future, moderators should consider the tradeoffs of using the think aloud protocol and if it is in their best interest, then they should review the instructions or best practices for conducting a think aloud study with a usability test. Since the papers that have been done comparing the protocols of Ericsson and Simon to Boren and Ramey, and they have found few differences in results for usability studies except that the participant was less confused; it would follow that usability professionals should consider the Boren and Ramey approach to put less stress on the participant.

The moderator should also consider when to use concurrent think aloud or retrospective think aloud. If any time on task metrics are important to the results then the retrospective think aloud protocol is a fair way of gathering addition verbalizations while not directly influencing the task. Considering that there have been many studies regarding retrospective verses concurrent think aloud and most conclude that they offer the same results and recommendations regardless of the practice used, it is worth using more often. Perhaps more research should be done regarding retrospective think aloud as it was not a focus in this report.

Finally, not only does the implementation of think aloud require careful consideration, but also the analysis is considered by many to be lacking. Ericsson and Simon put forth a complicated analysis process that would be impractical for most usability studies in industry. There is a need to develop a quick way to analyze the think aloud data. If protocols are followed during the session recording then there could be a set of protocols for analysis afterwards. With the amount of data collected just from the participant's verbalizations, there is a possible potential to learn a great deal more than is reviewed currently.

Conclusions

In summary, it has been shown that think aloud protocols are used quite frequently during usability testing to varying success. Some think aloud studies are successful because they are able to gather meaningful quotes from the participants. Others are successful because from someone’s comment they are able to interpret usability problems with the product. Some usability professionals will say that having the participant think aloud is unnatural and will skew the results. The problem is, when it comes to think aloud protocol and usability testing there is still a lot of research to be done on the subject. In many cases the theory and what is done in practice do not always coincide. The methods to analyze think aloud data are too time-consuming for the current pace of the industry. The need for usability testing is rapidly increasing with each new release on the market. Therefore, testing must be faster than ever before if there is any hope to keep up. The think aloud protocol is an effective tool for any usability practitioner when used properly, but it is still in need of more research that is practical in order to help usability studies in the future.

Appendix

Notes

1. **Moderator –** The moderator is the person who conducts the usability test by interacting with the participant. This can be an usability professional or an unbiased person who has been trained to moderate studies.

2. **Time on Task** – is a term referring to the literal amount of time a usability participant spends working on a finite task during a session. It is often hard to measure and usually measured in seconds.

3. **Session –** A usability session is defined as the time from start to finish with one participant in a usability study. Each study usually has at least five participants meaning, at least five sessions.

4. **Navigational study –** Is a type of usability study where the goal is to determine how the user perceives the navigation of the product, usually a website. Can the user get to a particular part of the product and know how to get back to the start easily?

5. **Exploratory study –** Is a type of usability study where there is no strictly defined goal. The idea of the study is to see how people will interact with the product and what features they are able find and use easily. It also helps gage people reaction to a new product or product change.

Bibliography

1. Boren, M. T. & Ramey, J. Thinking Aloud: Reconciling Theory and Practice, IEEE Transactions on Professional Communication, 43, 3 (2000), 261- 277.

2. Ericsson, K. A. & Simon, H. Protocol Analysis: Verbal Reports As Data, MIT Press, Cambridge, MA, 1984.

3. Krahmer E. & Ummelen N. Thinking About Thinking Aloud: A Comparison of Two Verbal Protocols For Usability Testing. IEEE Transactions on Professional Communication, 47, 2 (2004), 105 - 117.

4. Nielsen, Jakob. Usability Engineering. San Francisco, Calif. : Morgan Kaufmann Publishers, 1993.

5. Neilsen, Jakob. Useit.com: Jakob Neilsen's Website. http://www.useit.com

6. Norgaard, Mie & Hornbaek, Kasper. What do Usability Evaluators Do in Practice? An Explorative Study of Think-Aloud Testing. DIS 2006, June 26, 2006.

7. Pearrow, Mark. Web Site Usability Handbook. Rockland, Mass.: Charles River Media, 2000.

8. Pula, Katie and Sylvia Smith. Usability Testing Methods: Usability Methodologies and Strategies. McGill, March 15, 2004

9. Rosson, M. B. & Carroll, J. M. Usability Engineering: Scenario-Based Development of Human-Computer Interaction. San Francisco, CA. Morgan Kaufmann Publishers, 2002.

10. vaan de Haak, M. J. & de Jong, M. D. T. Exploring Two Methods of Usability Testing: Concurrent versus Retrospective Think Aloud Protocols. IEEE Professional Communication Conference Proceedings, Sept. 2003.