Usability II. Design of a User Study

August 1, 2009

Amalia Madalina Radoiu
amaliaradoiu@yahoo.com
Universität des Saarlandes
Matrikelnummer: 2510385

Seminar „Mensch-Computer-Interaktion“
Sommersemester 2009
Betreuer: Michael Kipp
Abstract. This paper talks about the steps as well as the different issues that should be considered when planning a usability study. The first step in planning such a study is to identify the goals of the study and choose between one of the two main approaches of how the data will be used, namely between formative and summative usability. Different purposes of evaluation are stated in the paper. When measuring the extent to which the user goals have been accomplished, performance and satisfaction are the main aspects that have to be considered. Further on metrics are introduced and ten different types of usability studies will be presented. There are also other important aspect to a study like budgets and timelines, participants, data collection, data cleanup and different evaluation methods such as lab tests, online studies and focus groups. Having planned a usability study properly can have important financial benefits for the project as well as considerably improve its success on the market.
1. Formative and Summative Study Goals

One of the main reasons for performing a usability study is to find out how one can improve the user interface. Developing a user interface with a high usability can not only increase the satisfaction of the user, but also improve the efficiency and speed of performing a task.

A product with high functionality is not enough if its user interface makes it difficult for most of the users to be able to make use of its functionalities. “Computer systems with poor user interfaces can have a financial cost.” (Stone D. (2005)

In the first step of a usability study one has to decide how the data collected during the study will be used in the development cycle of the product. The two main categories for classifying how the data will be used are formative and summative usability.

1.1 Formative Usability

Formative usability is part of a continuous development process; therefore it has an iterative manner. There are three important steps when using a formative approach, namely evaluating the product or design, identifying its shortcomings and making recommendations. The approach is being repeated until the aimed usability level has been achieved.

The formative approach helps identify the main usability issues of the product during its development cycle, the most common errors that the users are making, what works well and what not and it checks whether improvements have been made from one design iteration to another.

A formative usability study is useful only when we can impact the design during its development.

1.2 Summative Usability

In contrast to formative usability, the summative usability evaluates a product only after the development process has been completed. It checks whether a product or design has met its usability requirements and goals, it compares the product against other products and checks whether improvements have been made since the previous release.
Summative usability should include follow-up activities such as launching a new project to address certain usability issues or securing funding to enhance the usability of the present product.
2. Evaluation Purposes

There are four different types of evaluation approaches suggested by Rubin (1994): exploratory, validation, assessment and comparison. Each of the evaluation types is used at a different time in the product development cycle, according to the purpose of the evaluation.

2.1 Exploratory Evaluations

This type of evaluations is conducted in the first stages of the product development. It is a very informal way of evaluation, often conducted with low-fidelity prototypes and it includes having discussions with the typical users in order to establish the tasks that the users might want to accomplish.

Such evaluations are easy to arrange, since the claims are rather relative and not absolute. They are useful for exploring the user interface design features of the prototype, gathering feedback on preliminary designs and verifying assumptions held about the real users.

The data obtained from such an evaluation method is qualitative and it is used in establishing and evaluating usability requirements.

Such an evaluation method is both formative and diagnostic.

2.2 Validation Evaluation

Validation evaluations are performed toward the end of the development cycle or while the product is in use. It evaluates the product against a certain set of predefined usability standards. It can also be used to check how different components of the system work together.

In contrast to the previous evaluation method, here the claims that are made are absolute, that means a specific requirement is either fulfilled or not.

In order to be able to affirm whether a certain requirement has been met by the interface, we must
- establish a hypothesis, a testable claim or assertion that we are evaluating
- establish a null hypothesis, the possibility that the claim will not be met
- decide what sample size will support our assertions
- choose the participants randomly from the appropriate user population

This is a summative evaluation method.
2.3 Assessment Evaluation

Such an evaluation type is performed in the middle or at the beginning of the design development cycle, after the feedback obtained in the exploratory sessions has been incorporated in the design and the conceptual model has been created.

Assessment evaluation helps check whether the user’s tasks are supported by the design and identify the usability problems in the design. At this point the requirements of the design might need to be refined or even completely reformulated. This type of evaluation is usually formative and diagnostic, but it can also be used to establish measures of levels of usability, such as whether the user interface achieves its purposes for the best-case and worst-case task times.

2.4 Comparison Evaluation

Comparison evaluation can be performed at any point in the development cycle of the product. In the early stages of the life cycle it can be used to compare two designs with different interaction styles, whereas in the later stages it can be used to compare the efficiency of two alternative designs.

When comparing the evaluations, one has to take into account whether the differences in evaluation were influenced by the way we tested, by sampling or any other differences.

2.5 Evaluation Methods

The choice of the evaluation method depends on the number of participants and the metrics needed.

Lab tests require 4 – 8 participants and they are performed in one-to one sessions between the usability specialist and the participant. The participant receives a set of tasks to perform and the moderator notes down the behavior and the answers that the participant gives to the question. The most relevant metrics are issue frequency, type and severity. Self-reported metrics can be used although not always recommended.

Online studies involve more participants (10-50) users. Participants answer a set of questions and the data is collected automatically by a tool. Data is collected in a short time. Online studies are useful in collecting performance and self-reported metrics, but not of help for issue-based data.

Focus groups include 8 to 10 participants and are helpful in finding out the perceptions and attitudes of the users regarding a certain product. The only type of collected metrics are self-reported metrics.
3. Performance and Satisfaction as User Goals

Before planning a usability study it is necessary to know what the users are trying to achieve with the program and how they will use it. The frequency with which the program is used as well as whether it is used as part of the job or for entertainment are some of the issues that come into question.

The importance that the user gives to the efficiency while performing the task is also to be taken to regard during the planning of the usability study.

Nevertheless, the aesthetics of the design can also play a role for the users.

In the next two sections I am going to talk about the two important aspects when stating the user goals.

3.1 Performance

Performance is related to the level to which users can accomplish their tasks. There are certain metrics used in order to measure it such as the time needed by the user in order to accomplish the task, the effort (number of clicks, pages visited or cognitive effort) and the learnability level (that is the time needed for the user in order to become skilled in using the program). Performance is especially important for programs that the users need in their working field and upon their usage they have no choice.

3.2 Satisfaction

Satisfaction refers to the user’s impressions while interacting with the product. How the product met the user’s expectations, whether the user felt it was easy to use the product or he felt frustrated while interacting with it, are part of the satisfaction evaluation.

Unlike the previous metric, which was rather important when the user does not have a choice on whether he uses the program or not, this metric is a priority if the user has a choice on the usage of the program. This issue can be found in most websites, software applications and customer products.
4. Ten Types of Usability Studies

The type of metrics that we use when performing a usability study depends on the product that we are evaluating. Possible measurement criteria (Tyldesley, 1988) could be:

1. Time to complete task.
2. Percentage of task competed.
3. Percentage of task completed per unit time (speed metric).
4. Ratio of success to failures.
5. Time spent on errors.
6. Percentage or number of errors.
7. Percentage or number of competitors that do this better than current product.
8. Number of commands used.
9. Frequency of help or documentation use.
10. Time spend using help documentation.
11. Percentage of favorable to unfavorable user commands.
13. Number of runs of successes and of failures.
14. Number of times the interface mislead the user.
15. Number of good and bad features recalled by users.
16. Number of available commands not invoked.
17. Number of regressive behaviors.
18. Number of users preferring your system
19. Number of times users need to work around a problem.
20. Number of times the user is disrupted from a work task.
21. Number of times the user loses control of the system.
22. Number of times the user expresses frustration or satisfaction

There are several other types of metrics according to the purpose of the usability test or one can even define his own type of metrics. Ten different types of usability studies are going to be presented in the next sections.

4.1 Completing a Transaction

The purpose of some studies is to improve the usability of transactions such as purchasing or registering a product. The main metric while performing such studies is task success, that is whether the task could be completed or not.
In the case of a website, live website metrics such as the drop-off rate of a transaction are used. This helps identify the problem issues of a transaction.

Self-reported metrics can also be used. They include likelihood to return to the website and user expectations. Nevertheless efficiency is an important metric in this kind of usability studies and it is measured in units of time.

4.2 Comparing Products

Checking how the product performs in comparison to the previous releases may deliver important information. The metrics used depend on the products that we would like to compare. The most used type of metrics used in this kind of usability studies are combined and comparative metrics. Efficiency can be measured in the number of page views, time needed to complete the task or the number of steps taken. Task success is also used metric in this kind of studies.

4.3 Evaluating Frequent Use of the Same Product

Products that are used frequently such as kitchen appliances, DVD players or web applications need to be efficient and easy to use.

One of the most important metrics in performing such studies is task time, that is the time needed to complete the task. The number of steps of page views is one of the efficiency metrics used in this kind of testing. Learnability metrics are relevant in this kind of products, since the user is supposed to learn as fast as possible how to be efficient in using the product. Nevertheless self-reported metrics such as awareness and usefulness help identify the parts of the product that should be emphasized.

4.4 Evaluating Navigation and/or Information Architecture

Often usability studies concentrate on improving the navigation and/or information architecture. Such studies use wire-frames and or partially functional prototypes. A wireframe is a basic visual guide used in interface design to suggest the structure of an interface and relationship between its pages.

One of the relevant metrics used in this kind of studies is task success. Lostness is another metric for evaluating navigation and it refers to the number of steps the user had to take in order to complete the task.

In order to understand how the participants organize the information one uses card sorting. This approach is especially useful for defining web-structures and it works the following way: Participants are given cards with names of items that should be categorized. They are asked to group them in a way that makes sense to them and may also be asked to give names to the groups they have created.
4.5 Increasing Awareness

There are design studies aimed at increasing the awareness of a specific piece of function or functionality. Online advertisements are an example of where one would use such a usability study. Monitoring the number of interactions with the element we want to increase the awareness for can confirm the awareness but not demonstrate the lack of awareness. Sometimes self-reported metrics are used in order to see whether the participants are aware of a specific design element.

Memory is a self-reported metric which is relevant for this kind of usability tests. Participants are shown different elements of the design and asked whether they have seen them during the task.

With the help of technology one can use behavioral and physiological metrics such as eye-tracking data. Live website data is a metric which allows us to check how the traffic patterns changes between different design patterns and identify the awareness to specific elements of design.

4.6 Problem Discovery

The purpose of some studies is to identify major usability issues. In such studies participants may also generate their own tasks.

Issue-based metrics are good way to identify problems. Such metrics are chosen accordingly to the functionality of the product for which we want to measure the usability level or performance. By analyzing the frequency and severity of the shortcomings one can find out how many of them are observable to the users.

4.7 Maximizing Usability for a Critical Product

Designs which can have a critical impact, such as life-saving applications need to be efficient. Shortcomings in the usability of such products can result in a critical negative outcome. The user performance is measured against a target goal. If the critical product does not meet its target it should be redesigned. Such studies should involve a large number of participants.

Errors (the number of errors made by the user while performing a certain task) are one of the main metrics to this type of studies. Task success plays an important role in evaluating critical products. Since efficiency is one of the goals of the product, this metric is also part of the evaluation measurements.
4.8 Creating an Overall Positive User Experience

The overall positive user experience is especially relevant for products where the user experience has an important role, like for entertainment applications. The opinions and feelings of the user while interacting with the product ponderous that the performance itself. Satisfaction, which is a self-reported metric is one of the key metrics in such evaluations. The user experience is supposed to surpass the user expectations in order for the product to be successful.

Apart from the self-reported metrics, physiological metrics can also be used such as pupil changes, sweating etc.

4.9 Evaluating the Impact of Subtle Changes

There are design changes whose impact on the user behavior is not obvious. Such subtle changes can be the font type and size, position, color etc. Subtle changes can also be non-visual like content or terminology changes.

Live-site metrics from A/B tests are a key measurement approach when evaluating the changes of such metrics. A/B testing refers to comparing the control design against an alternative design. An example would be diverting a part of the traffic to a website to the alternative design and the other part to the actual design and afterwards compare the metrics results of the two different design variants.

4.10 Comparing Alternative Designs

The alternative designs are being compared in the early stages of the product development. Prototypes suggested by different teams are evaluated against a predefined set of metrics and then the results are being compared.

There are two solutions in order to avoid the learning effect while performing such studies. The first one would be that each participant sees only one prototype. The second solution would be to counterbalance the design. By counterbalancing the design we separate different groups of users and each group receives the design prototypes in a different order. This way we can avoid the learning effects.

Issues-based metrics are good approach to decide which design has a higher usability. By using self-reported metrics one can find out from each participant which of the different prototypes they prefer.
5. Other Study Aspects

5.1 Budget and Time

The time and the costs of a usability study depends on the evaluation method that is used.

In a formative study with a small number of participants (ten or fewer), collecting the metrics has little impact on the budget and time, if at all. Gathering the data and analyzing it should take no longer than a couple of hours.

Lab tests have a more significant impact on the budget and time since the number of users is higher (more than 10). Recruiting and compensating participants is the most significant cost. The time needed for running the tests with each participant has the biggest impact on the timeline. Time is also needed in order to analyze the data.

Online studies are usually performed on a large number of participants. The main costs of such studies are related to paying the person developing the study and analyzing the data. Half of the time is spent setting up the study and the other half is spent on cleaning up and analyzing it. Little to no time is used for collecting data.

5.2 Participants

The criteria used for recruiting participants should be well-defined in order to avoid recruiting participants from the wrong category of users. Some of the used criteria are the experience with the web, age and experience with financial transactions. One can take different participant types such as old and new users.

The number of participants needed is influenced by aspects like goals of the study, complexity of the product or diversity of user population.

Formative studies need 6 to 8 participants and by distinct groups at least 4 from each group. Summative usability studies require 50 to 100 representative users, but could also be 20. If the aim of the study is to test potentially subtle design changes one needs at least 100 participants.

After deciding on the number and categories of participants one has to choose a recruiting strategy. There are four different approaches to doing this. One can make the selection by generating list of possible participants from customer data, sending requests via email distribution lists, refer to a third party to handle recruiting or make announcements on the web or emailing a specific group.
5.4 Data Collection from the Users

An essential part of the user-centered design is to observe the users in the natural environment in which they are going to use the product, for example their workplace, home. This can provide the evaluator with information about what the users do as well as what they like and dislike about the product. This can be done both directly and indirectly.

Direct observation can be performed in many ways, but they are usually classified as either field studies or controlled studies.

Field studies are done by observing the users in their natural home or work setting and making notes about interesting behavior of the user.

Controlled studies are performed in a setting other than the usual one of the users, for example in a usability laboratory. In this case the user is given certain tasks that he has to perform, while the observer records the time the user needed in order to perform the task or certain particularities regarding the way the task was performed.

Direct observation is limited by the fact that it is impossible to gather a full record of the user’s action in an observation session.

Video recording provides a permanent record of the indirect observation sessions. This is considered to be more objective than direct observation.

One disadvantage of this method would be that there is a lot of work to do after the indirect observation session, in order to analyze the collected data. Also one has to try to reduce the impact that the video equipment has on the behavior of the users.

Interviewing the users involves talking to or questioning the users and has to be planned in advance. There are two types of interviews: structured and unstructured. A structured interview has predefined questions asked in a set way, whereas an unstructured interview has some set topics but no exact sequence.

Questionnaires and surveys are an alternative to the interviews and are used for gathering more precise information. A questionnaire can contain closed or open questions. In closed questions the answers that the user can give are either yes, no or I don’t know or it can be a multipoint rating scale called a semantic differential. An open question gives the users the opportunity to freely answer what they like.

5.5 Data Cleanup

Data Cleanup is necessary in order to be able to perform the analysis. In the following, the main steps of data cleanup are being stated.

Filtering data - extreme values such as task completion times should be removed
Creating new variables – building on a raw data set or aggregating data
Verifying responses – verify user responses, check the frequency of wrong answers
Checking consistency – make sure the data is captured properly
Transferring data – transferring the data to another program to run statistics
5.6 Conclusion

There are many aspects that need to be taken into consideration when planning a usability study: evaluation method, user goals, metrics, test participants, budget and time, data collecting and cleanup. One needs to carefully analyze all these factors which are part of the usability study. If the evaluation method, metrics, number and categories of users as well as the way the data is being collected and analyzed are chosen properly, it can be very helpful in getting important information on how the usability can be improved and what the shortcomings of the products are. On the contrary, having taken the wrong decisions regarding one or more of the aspects that come into consideration when planning a usability study can result in wasted money and time as well as probably rather decreasing the usability instead of improving it.

All this being said, the design of a usability study is an important part in assuring the success of the product.
6. Bibliography


T. Tullis, B. Albert (2008) Measuring the user experience, Chapter 3: Planning a Usability Study
