

**Usability Engineering**  
**Dr. Debayan Dhar**  
**Department of Design**  
**Indian Institute of Technology, Guwahati**

**Module - 12**  
**Lecture - 36**  
**Usability Testing**

Welcome to module 12 lecture number 36, we are at the fag end of this course, the last module. And in this module, we are going to discuss about Usability Testing of the product that you have developed.

See until now we have focused on the various aspects of how designers identify needs, define the requirements, establish benchmarks, focuses on the brief refers to the persona, scenario based on the user studies they have conducted. And then they start working towards developing concepts which would probably address the requirements that they have identified.

Now, in this journey if you see one of the vital aspects of ensuring that the product or the software that you have developed works is to see and ensure how your users perform or experience the product while they use the product. And it is in this regard that usability testing plays a major role.

Now, while we discuss about usability testing you must understand the fact that the concept of testing of usability parameters or usability as a parameter is focused on identifying and testing basic constructs that leads to the concept of usability. And, how do we do that? We do that by ensuring various methods and processes through which we capture the experience of the user of your end user while they are working with your product or they are completing a task with your product or and also, they have completed the task.

So, in both these activities or, this timeline we use specific techniques to capture the data to extract the data that provides us with the insight about the idea of usability about the product that we have developed or conceptualized.

(Refer Slide Time: 03:16)

## Usability Testing

- There are generally two types of usability tests: finding and fixing usability problems (formative tests) and describing the usability of an application using metrics (summative tests).
- formative—providing immediate feedback to improve learning vs.
- summative—evaluating what was learned



Now, to begin understanding usability testing we must first understand that while we started discussing about the design user center design approach we did talk about the various aspects of the linear process. While, discussion that we had if you remember we did talk about that it might look like a linear process, but it is not so, it is a iterative process.

And why it is iterative? Because every now and then when there is a state of confusion in the mind of the design team they fall back to their most relied and time-tested partners who are the users here. And get feedback from them based on their inputs in order to clear the state of confusion for the team of design.

Now, from that perspective if we see usability testing then we can classify usability testing primarily into two different categories or two different types. And these are the formative usability testing and the summative usability testing right, the formative usability testing and the summative one. And these are the generally two types of usability tests that are conducted in the life cycle of the design process that we have been discussing about.

Now, in formative test what happens as you run through your concepts as you start detailing about concepts you fall back to your users whenever there is an issue whenever you are unable to, you mean as a designer. You are unable to figure out certain conditions, certain states, certain actions or even certain features you there is a state of confusion in your mind which one

to decide which one to go for which one to utilize which one to select in those situations formative tests are conducted.

And they help you in finding and fixing the usability problems in your concept that you have generated. While, in summative tests it describes the overall usability idea or the parameter of the application in terms of certain metrics that we will discuss. Now, what is this idea of formative and summative?

See formative means as something is going while the product is being designed, while a product is being conceptualized tailed out certain situations or certain states are prototyped and then quickly they are being exposed to the users for their feedback.

Now, this is an ongoing process of development is happening, but a summative usability testing is not like that it does not focus or the focus is not while you are developing the product you go back to the users. No, once the prototype is ready completely ready and you are in position to run the product the application you go to your users get their data and see and compare how it has performed with the old version or the new one.

(Refer Slide Time: 06:59)

### Usability Testing

- The bulk of usability testing is formative. It is often a small-sample qualitative activity where the data take the form of problem descriptions and design recommendations.
- two types of summative tests: Benchmark and Comparative. The goal of a Benchmark Usability Test is to describe how usable an application is relative to a set of benchmark goals. Benchmark tests provide input on what to fix in an interface and also provide an essential baseline for the comparison of post-design changes.

Now, the bulk of you know usability testing is formative we talked about cognitive work through approach. And it is often a small sample small number of people or your users who participate in your study which we call as samples, qualitatively interact with the product in terms of they give you qualitative feedback. And these are activities that are performed in terms of process in terms of approach what is happening inside them.

So, it is often a small sample qualitative activity where the data take the form of problem descriptions and design recommendations. Now, the major two types of usability testing which is primarily summative in nature are carried out with an objective to benchmark and to compare. See benchmark is of highly important for us, is not it? We discussed in our earlier modules how benchmarking allows the design team to identify the parameters the contracts based on which they are going to conceptualize.

The values which they need to break or with which they need to reach in order to ensure that the product the software that they have developed sustains the competition in the market. And similarly, comparison is also important. Comparison is important to quantitatively identify the incremental differences that your new design concepts, has brought in if you are working in a redesigned project.

So, you compare your new design ideas with the old design interfaces and see how incremental innovation has happened. So, that the metrics that you are going to identify they would let you know whether the difference exist or not or whether it is much worse than what it was previously. So, the goal of you know the benchmark and the comparative one is to describe how usable an application is relative to a set of benchmark goals.

And the benchmark tests provide input on what to fix in an interface and also provide an essential balance for the comparison of the post design changes.

(Refer Slide Time: 09:36)

## Usability Testing

- A Comparative Usability Test, as the name suggests, involves more than one application. This can be a comparison of a current with a prior version of a product or comparison of competing products. In comparative tests, the same users can attempt tasks on all products (within-subjects design) or different sets of users can work with each product (between-subjects design).



And a comparative usability test it involves comparison either between the old application that which you are redesigning or a comparison across your competitors. Now, this can be a comparison of a current with a prior version which I said or even of the comparing products.

So, in comparative tests the same users can attempt tasks on all products. So, the same tasks are being compared across the various version of the products across the various competitors and the design team actually see which one is better in comparison to the other right. And the different sets of users who can work on each product.

See when we talk about this comparative usability test one of the important issues that are often discussed among the research team is to whether use the same people across different platforms. So, what you say? You have 5 people and you are using the same number of people, the same people. Not only the number the same people the same people are involved in conducting a task in platform 1, the same people are also involved in completing the task in platform 2, platform 3 so on and so forth.

Now, the idea is you are using the same samples and what is the main idea behind using the same samples? The idea is that if you are using same samples they would probably in a position to compare the tasks across the platform. Because they have used a different they would be exposed to different design features of the same task and they would be in a much better

position to explore compare these features and provide you with data.

And when you are using the same samples across multiple platforms conducting the same task we call that as a within subject design. Subject means here subjects means samples. Now, there is also a line of thought which says that if you are using the same samples over and over again across different products, across different platforms then there is a learning effect.

That means, the task is there he is he probably he probably he or she probably knows a task in a much deeper way. And there would be a learning effect across these platforms and that might influence their responses while they are comparing the platforms. Now, in that line of thought this is also refers to biases, the individual biases that might influence these study results. In those case, researchers has come up with a different approach which is called as the between subject design.

If you see in this slide we talked about between subject design. Now, what is a between subject design? Now, between subject design means you are not using the same samples over and over again across the different application platforms. What you are doing is you are using the same number of peoples; I mean, the number is same, but they are not same across all platforms you use different samples for the different platforms.

But keeping in mind that the samples are off or they have similar individual characteristics. For example, if in platform a you are using 5 samples who are of male and in platform b if you are using 5 samples who are female then they are not comparable because gender differences exist. And because of gender we assume that individual characteristics will also change which might influence the study results.

So, ideally if we say in group a we have around 20 30 percent around out of 5 samples 2 are females and 3 are males and in for platform b we have around 2 males and 3 females it is some extent comparable.

So, if you are conducting a between subject design you are addressing the concept of bias, you are addressing the concept of learning influences. And thereby you are ensuring that the data

that you are getting from your samples based on the metrics that you have identified they can be compared and a true figure can be identified and a true outcome can be generated.


The only important aspect here for you to keep in mind is to ensure that the samples are comparable they do not differ in terms of their individual characteristics, so much so that their data's can be questionable. This is the concept of within group design or within subject design and between subject design. And these are very important concepts that you must keep in mind while you are planning for usability tests.

(Refer Slide Time: 15:57)

**Usability Testing**

**Metrics**

- Completion rates: Also called success rates typically collected as a binary measure of task success (coded as a 1) or task failure (coded as 0). report completion rates on a task by dividing the number of users who successfully complete the task by the total number who attempted it. For example, if eight out of ten users complete a task successfully, the completion rate is 0.8 and usually reported as 80%. You can also subtract the completion rate from 100% and report a failure rate of 20%.

 Dr. Debasis Sahoo  
Department of Design

Now, some of the metrics that are extensively used in usability testing few we have discussed earlier in the initial lectures while we will talk about some of them in detail in this lecture. The first most important metrics is the completion rates now, what are completion rates? The word itself would tell you what this means by completion.

What does your user complete see your interface is a medium to complete what? That is the question, you must ask you are designing a software you are designing an application. But the goal of your user is not to use the application just because they want to use it they want to reach a goal, they want to complete a task and that is their central objective.

So, a completion rate is about how successfully your samples or end users are able to reach their goal or complete the task. So, completion rates are also called success rates in many papers, in many case studies of usability or in the domain of human computer interaction user experience design you would see people talking about success rates. These are same, success rate means completion rates and these are typically collected as a binary measure of the task success. I mean, when we talk about completion rates or success rates what are the possibilities that can happen?

There are only two possibilities that can happen either there is a task success, which is coded as 1 or there is a task failure which is coded as 0. Now, task success means what? Your end user is able to complete the task and therefore, you have identified that if a person is able to complete the task make it 1 he has his score is 1.

And if it is if he is unable to reach to the goal if he is unable to complete the task then the code is 0. So, these are reported completion rates on a task the report completion rates on a task by dividing the number of users who successfully complete the task by the total number who attempted it. And that what will give you the percentage of success rates or completion rates of that particular task using that interface.

So, some of the examples like you know 8 out of 10 users complete a task successfully. So, the completion rate is what 80 percent right. So, you can also subtract the completion rate from 100 percent and report a failure rate. So, the failure rate is 20 percent. Now, say for example, your stakeholder who has already a new software who has already a software rather he says that in this software the failure rate is around 40 percent I want you to reduce down to 10 percent.

Now, that is very crucial, now he has already given you a benchmark and the benchmark is what? Identify the usability issues identify why they are unable to complete or unable to the success rate is less it is around 60 percent. And with your new design you are supposed to improve it to at least to 90 percent right. And that is what which we understand by completion rates or success rates.

(Refer Slide Time: 19:47)



## Usability Testing

### Metrics

- **Task Time:** Task time is how long a user spends on an activity. It is most often the amount of time it takes users to successfully complete a predefined task scenario but can be total time on a webpage or call length. There are many ways of measuring and analyzing task duration:
  1. **Task completion time:** Time of users who completed the task successfully.
  2. **Time till failure:** Time on task until users give up or complete the task incorrectly.
  3. **Total time on task:** The total duration of time users are spending on a task.



The next important metric is about task time now, what is task time? Now, task time is how long a user spends on an activity it is the temporal dimension, the time dimension. It is most often the amount of time it takes users to successfully complete a predefined task scenario, but can be total time on a web page or it can be called as a length. And there are many ways of measuring and analysing task duration.

Now, the various ways through which it is done it is by these three metrics task completion, time till failure, total time on task. Task completion time means time of users who completed the task successfully code 1. What is the time?

Time till failure time on task until users give up or complete the task incorrectly that is time till failure. Total time on task: the total duration of time users is spending on a task this is irrespective of the completion time or the failure time. Now, understand this situation why it is so important.

See in the, these era of internet in this era of web the enigma of an interface to hold its users earns revenue. So, the more time you ensure that your users are using your interface holding on to your services, getting into the task, completing it and still holding on to those products ensures that you earn return on investments, ROI right.

And therefore, all designers eye that sweet position of identifying or using ways through which they can hook their users in ensuring that they spend most of the time using that product alone. And therefore, you would see there are so many products in the web they have started using extensions, plug-ins to ensure that they do not leave their space, but while working in that space if they want something else they can use those services and still remain in their applications.

So, time is money when the focus is on the web or the internet. And the parameters that will help you understand how your users are playing with time are the task completion time, the concept of task completion time, time till failure and total time on task. And here the key parameters that would make you analyze the total time on task or task completion time or time till failure in a more appropriate way is your ability to gauge or understand the individual differences among your users.

Whether your user is a novice one, whether is your user, is an expert one or whether he is an intermediate one. Because you know that if he is a no novice one remembers to what we have discussed earlier. The mental operator is going to use this much more in comparison to an expert user for whom the task has been has become a routinely activity. And therefore, the mental operators that he is going to use are far comparatively less than the novice users.

So, in this way if you look at these parameters and metrics you would be able to identify more informed decision. Or, you would be able to take more informed decision regarding what needs to be taken considered or what needs to be changed or addressed in the concepts that you have come up with.

(Refer Slide Time: 24:06)

## Usability Testing

### Metrics

- Errors: Errors are any unintended action, slip, mistake, or omission a user makes while attempting a task. Error counts can go from 0 (no errors) to technically infinity (although it is rare to record more than 20 or so in one task in a usability test). Errors provide excellent diagnostic information on why users are failing tasks and, where possible, are mapped to UI problems. Errors can also be analyzed as binary measures: the user encountered an error (1 = yes) or (0 = no).



The next metric that is of a major concern, because you remember we talked about heuristic evaluation and there also it is a major concern is errors, errors that your end users commit. Now, errors are any unintended action, these are unintended action, these are slips, mistakes, these are omissions a user makes while attempting a task. Now, error counts can go from 0; that means, there is no errors while the user has completed the task to technically infinity infinite errors ok.

But research says that it is very rare to find more than you know to record more than 20 errors in a particular task or in a usability test. So, errors provide excellent diagnostic information on why users are failing tasks and where possible are mapped to the UI problems, the interface related problems the way it is being structured, hierarchy, classification, naming. So, errors can also be analyzed as binary measures and these are the user encountered an error; that means, the state is called 1 yes or no that is 0.

(Refer Slide Time: 25:34)

## Usability Testing

### Metrics

- Satisfaction Ratings: Questionnaires that measure the perception of the ease of use of a system can be completed immediately after a task (post-task questionnaires), at the end of a usability session (post-test questionnaires), or outside of a usability test. Although you can write your own questions for assessing perceived ease of use, your results will likely be more reliable if you use one of the currently available standardized questionnaires (Sauro and Lewis, 2009).



The next one which is a subjective score is of primary importance in the usability testing situation and this metric is about satisfaction ratings. So, question is that measure the perception of ease of use see remember here we are focusing on the perception of use ease of use of a system can be completed immediately after a task, post task questionnaires are used here, at the end of the usability session or outside of the usability test.

Although, you can write your own questions for assessing perceived ease of use your results will likely be more reliable if you use one of the currently available standardized questionnaires like the questionnaire of Sauro and Lewis of 2009.

This is just a for a reference I have given. Now, when we talk about satisfaction ratings what is happening in the mind of the user let us just understand the situation. So, you have given this interface the software you have asked your user to complete the task. The moment your user hears about the task he will have a mental model right, we discussed about this earlier also.

And while he starts working towards the task using your interface starting interactivity his mental model is influencing the way he is searching for information, he is figuring out what needs to be done, he is comprehending what is being presented in front of his screen all these are has direct influence of the mental model that he has. Now, once he completes the task what happens?

In real time there is a comparison between what he had in his mind and what is being given to him. If there is a disconnect between what you as a designer has designed for him and what he what he wanted then this subjective ratings like satisfaction ratings will go down. Because he will evaluate this difference and then probably he will give you the score ok, I think this is not very nice, I did not like the interface.

Now, there is always a frame of reference now when he says that he did not like the interface or he did not like the way the task is was being framed. He has a benchmark he has a frame of reference and using he is using that frame of reference which is his mental model to evaluate this satisfaction ratings. So, therefore, the role of mental model and the role of your user study becomes so critical that if you get things wrong to understand the mental model of your user your concepts will fall flat they will not work.

And coming down to the satisfaction ratings how do you capture these ratings? You capture these ratings based on the task that your users have completed and then you go to them you ask questions you can give a question here. And then based on that questionnaire he or she provides you the answer. Now, there are critical issues about designing a questionnaire also.

So, question is what? When we talk about a questionnaire a question is ideal in instrument right. You have seen instruments being used in labs, in physics, in chemistry there are so many instruments that measure a certain value and allow you to interpret the data, interpret the situation. Similarly, for a user experience designer or human computer interaction designer the instruments are the questionnaires.

And if these questionnaires are not are incorrect then they would not give you perfect values of these metrics like satisfaction ratings. And for this questionnaire to be perfect what is important for you to understand is that these questionnaires need to be valid and reliable. So, the concept of validity and reliability comes into play.

Now, when I say valid what do I mean by that? I mean by that if you are, if your intention if the intention of the research is to measure satisfaction score he must design an instrument, a questionnaire such that it measures only satisfaction. It should not measure something else, it should not measure some other parameters for example, his future plans or how does he

envisage a situation in the future or what was his past experience so on and so forth.

Now, if it measures all those parameters except satisfaction then the questionnaire then the instrument is not valid. And there are very ways through which we test for validity also there are some statistical way there are some subjective way through which validity is measured. Now, similarly along with the concept of validity the concept of reliability is important your instrument needs to work every time the same way.

If it is measuring satisfaction it must measure satisfaction here while you are working it should business satisfaction in a different situation also. Consider the situation that you are wearing a watch and you are trying to fly from New Delhi to New York and you are not changing your shifting or time of your watch. Now, will your watch or the time that you would see if you land in New York be reliable?

It is still valid, because it is giving you time, but then it will not be reliable because it will not show you the New York time, it would show you New Delhi 's time that is the concept of reliability. Now, at this stage while you are being exposed to the concept of usability it is not likely that you would embark on a situation on a questionnaire design.

So, therefore, in order to reduce your loads or your tasks one of the ideal ways through which you can focus on having in questionnaire or an instrument that is ideal is to take questionnaires that are already being tested throughout the world and has performed exceedingly well in terms of their validity and reliability. And one of these questionnaires that I have just mentioned is the Sauro and Lewis questionnaire of 2009, where they measure the satisfaction ratings.

You can use those questionnaires and after the task completion you can give it to your samples and they can provide you with the data.

(Refer Slide Time: 33:26)

## Usability Testing

### A/B Testing

A/B testing, also called split-half testing, is a popular method for comparing alternate designs on web pages. Popularized by Amazon, users randomly work with one of the two deployed design alternatives. The difference in design can be as subtle as different words on a button or a different product image, or can involve entirely different page layouts and product information.



Now, these are some of the metrics, now we would talk about some of the ways through which usability testing is conducted. One of the most celebrated known way of conducting usability testing is called the A/B testing often it is also called as the split half testing.

And it is a very popular methodology across corporates, organizations, researchers in the field of HCI and design, UX design. And the idea behind this is they compare alternate designs. So, when we talk about A and B we are essentially talking about two states of the same of a product state A and state B, design A and design B.

And it was ideally popularized by Amazon and users randomly work with one of the two deployed design alternatives. And the difference in design can be as subtle as different words in a button or a different product image or different architecture, different ways through which information is grouped different way of navigation so on and so forth.

Many interface features can be can get differed in terms of design A and design B. And while in real time they are given to different samples the data are extracted in terms of their task completion rates, the errors that they commit, the satisfaction ratings and an insight is being generated based on the performance of both the screens. So, this is about A/B testing or the split half testing.