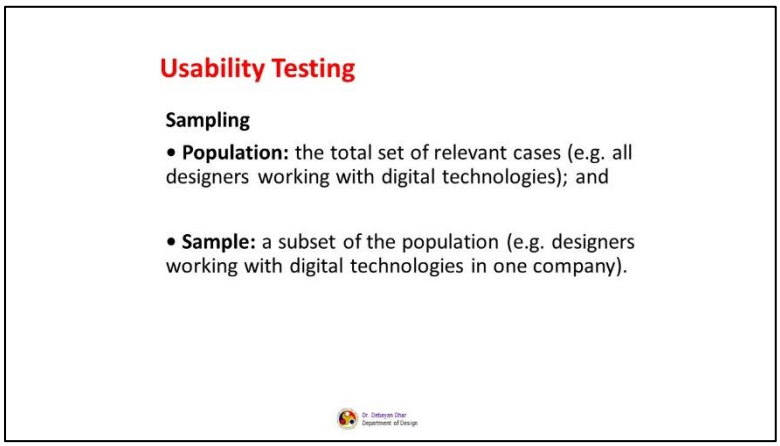
Usability Engineering
Dr. Debayan Dhar
Department of Design
Indian Institute of Technology, Guwahati

Module - 12
Lecture - 36
Usability Testing

(Refer Slide Time: 00:32)



We will now discuss about sampling. Now, it is important to highlight here that we are conducting experiments in order to understand how the new designs or the concepts that you have created performs against a standard or the old design. And the kind of testing we do for that which we have discussed in the split half of the heavy testing; one of the important aspects for conducting this test is conducting users, is recruiting users.

Now, users are often referred to as samples in the structure of experimental design. These are the people who are going to use our product; they can provide us with their data's as they use the product, the specific matrices that we have talked about. And based on the comparison, we would be able to realize whether the design considerations that we have made really performs or does exhibit the properties that we want from them.

Now, once we are done with this kind of structure, we know that ok this is the new concept that we are going to test, this is the old one or we can have various concepts among which we want to compare. The important aspect here is how you recruit your samples or how do you conduct this study in order to ascertain the assumptions, the hypothesis that you are going to test.

We would come to that later; right now, we will focus about the concept of sampling, the concept of population, parameter and the sample parameter. Now, when we talk about population, we mean the total set of relevant cases.

For example, if you have designed a software, a product for people who intend to book flight tickets or for those who are into online learning. In these cases, population means the entire population of people who want to buy tickets or who are into online learning. So, the entire subset or the total set of those relevant users are considered to be as the population.

Now, you need to understand that the intention of conducting an experimental study here is to understand how our designs would perform when they are being offered to the populations, right that is what our intention is. And how do we study that? Is it possible for us to go to each person, which is the total set that we are referring to and ask them to use our product and then provide us with their insights or capturing their matrices to understand how these designs are performing? It is not possible.

So, in that case what do we do? We identify a smaller set of population. So, we recruit few people from this population and on these, on this smaller number of populations, we conduct our experiments. And the intention is that, this smaller number of people who has been extracted from the population; if we conduct a study with them, we would be able to gauge how our population will perform.

This smaller set of population that we are extracted are referred to as samples, ok. So, that is the difference between sample and population. So, sample is a subset of the population; we are extracting few of our relevant cases from the total population parameter and we are calling them samples. And based on this study on the samples, we would be able to extrapolate about

how our designs are going to perform when they are exposed to the population.

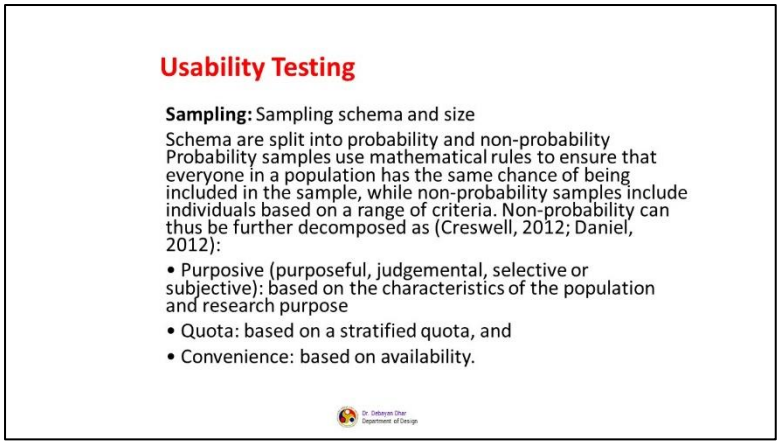
So, in the slide we have talked about examples of designers working with all digital technologies; if that is what our population is, then we would recruit few of them, create a group or individual people and then that becomes our sample for this study. Now, here it is important for us to understand that the primary objective of conducting these experiments is to ascertain the assumption that we are going to make.

And what is this assumption? It might be that the design that we are proposing would influence the satisfaction of our samples, of our population; that means we hypothize that if we make these changes in this interface or in this software structure in this way or if these are the changes that are being done, then the satisfaction level of our samples would increases, satisfaction level of our population which is the final objective would be higher than what they are currently now.

It can also be in terms of errors they commit, we can assume that the errors would be reduced if this kind of interfaces are used right; time on task would be much less in comparison to what is existing or the other concepts that we have that we have come up with. So, that is what we often refer to this when we talk about the assumptions that we are discussing about, these assumptions about; these assumptions are nothing but predictions.

And this predictions about how the interface or how the design would perform in terms of the metrices are being tested by conducting an experiment, where we recruit sample from the population concerned to establish whether our assumptions or the predictions that we have made holds its ground or they are untrue.

(Refer Slide Time: 08:19)



Now, there are different ways through which samples are recruited and that is what we call a sampling. Sampling means we are talking about sampling scheme and size. Now, schema, the sampling schema that we are referring to are split into; these are types of ways through which sample is recruited into probability and non-probability concepts or ways.

Now, probable probability samples are those that use mathematical rules to ensure that every one in a population has the same chance of being included in the sample, while nonprobability samples include individuals based on a range of criteria. And these nonprobability samples can be farther classified into what you can see in this slide as purposive, quota and convenience. Now, it might look little bit alien to all of you about why there is so much discussions about how we are recruiting our samples. Your concerns may be right to some extent; but let me explain to you the situation why do we need to follow this kind of practices to ascertain that we recruit correct samples. See when we are conducting experiments as the researcher or the design researcher; you as a person influence the way, you as a person can influence the way the experiment is conducted.

And how? Consider the situation that, you have initiated an experiment where you want to see how a particular design would perform in terms of time on task, error, satisfaction course. And what you have done is, you have recruited all your friends to conduct this study; because you

know them, they can be approached very easily and because you have a friendship with them, they cannot say no to you, you know because you have requested a favour from them.

Now, what has primarily been observed in this kind of situations is that, the personal relationship that the researcher has with the samples; influences the way the sample responds to the questions, to the questionnaires. And therefore, the data that are in concerning, the data that are in concerned are often considered as biased data.

Now, for example, the situation is that I call my very good friend to come and test an interface that I have designed and it might be a case that even if he does not like the interface, just to ensure that his response does not hurt you; he may say or she may say that oh it is absolutely wonderful, I feel it is very nicely done, while the actual reasons or the actual situation might be very very different, ok.

So, these kinds of individual biases can influence your study results. To ensure that we do not conduct or get influenced or our study does not get influenced in any such conditions; what we do is that, we follow standard sampling schemas.

Now, it is also true that in certain cases you need to identify people through whom you know, with whom you have an understanding and we will discuss about that. And for these kinds of situations, generally scientists and researchers follow these two types of sampling schemas; one is the probability sampling, the other one is the non-probability sampling.

Now, in a probability sampling schema the idea is say, if you have a population parameter of hundred people; say my population is I am working in a study or doing an experiment that, concerns the population of a particular college, say for example, any IITs or NITs or any other colleges. And in that college, the total number of populations of that particular group is around say hundred people.

Now, for each one of them, each one of them has a chance to get recruited in your experiment and that is the concept of probability, ok. So, random number tables, other ways through which

you know you first of all identify those samples are used to ensure that, randomly you select people and every one of them has a chance to get recruited for the experiment.

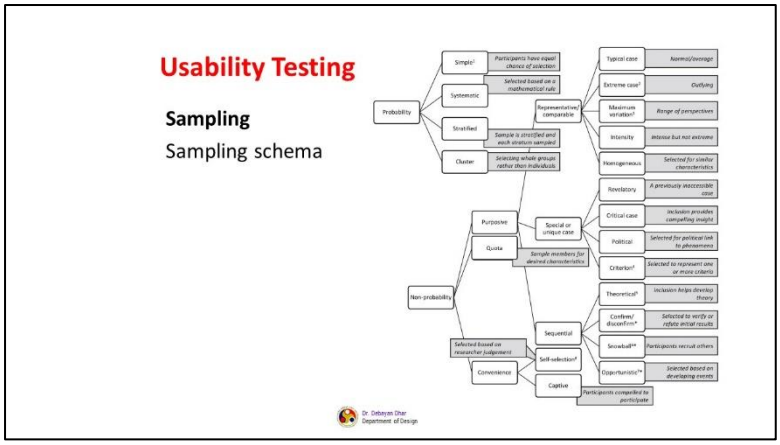
Now, that is also another type of sampling schema, which is called as the nonprobability sampling. And the types are this purposive, quota, convenience. Now, in purposive nonprobability sampling, you purposefully with your own judgments you select the subject and these are based on characteristics of the population and research purpose. For example, say we are conducting a study on people who has partial blindness and we want to test how the interface are working on with them.

Then we have to purposefully select people who suffers, who are partially blind in nature; that is an example of purposive sampling. We can also have quota samplings, various quotas; for example, gender can be a quota, we can have people whose income range comes from a particular group that is also a quota sampling. And there can be stratus based on which these samples can be extracted.

For example, we want to focus on male gender above the age of 35, below 60, who are working in software's for our samples, that is a strata, stratified sampling; it is a quota based on quota, but there are various stratus.
The first strata are that it needs to be a male gender; the second strata is the age one, where we are talking about more than 35, but less than 60; the next strata is we are focusing on people who only work in software. There can be other strata that people who are product managers only we are focusing on that. So, this is quota sampling. The other one which is called convenient sampling is generally used in health cares.

For example, you know you want to see how a particular treatment affects a particular group of people who are suffering from that. So, it is not possible for the doctors or the researchers in healthcare innovation to go and conduct this study randomly. So, what do they do; based on the availability, based on the patients who are coming for treatment, they use those as samples for doing their experiments, that is called convenient sampling.

(Refer Slide Time: 16:34)

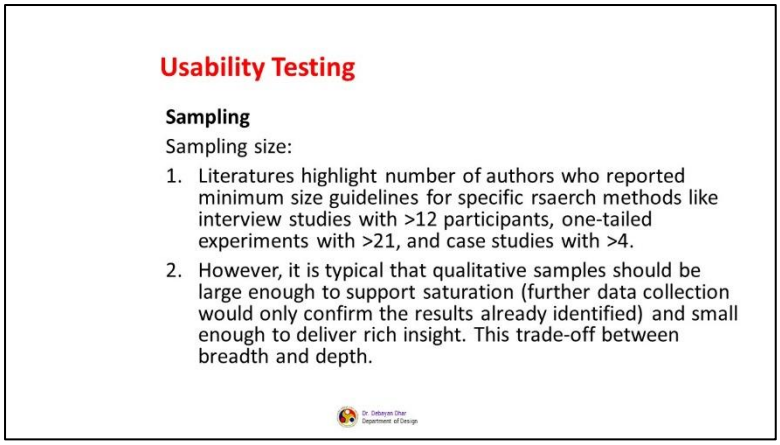


Now, what you seen in this slide here is a detailed structure of the sampling schema. You see, I have classified the two types probability and non-probability and internal types also in this image; you see in probability we can have simple probability samplings, we can have systematic probability samplings, stratified, cluster also. So, sampled is stratified and each stratum sample. So, this also stratified in random sampling in probability sampling.

So, you create these strata and from that strata, you randomly select. Systematic sampling means selection based on a mathematical rule, probably you can use a random number table, identify and give numbers to particular people and randomly based on the random number table you can select those participants, right. Then you have non-probability sampling, where we talked about the purposive and the quota sampling and the convenient sampling, right.

So, these slides provide you with a detailed structure, which you can refer to in terms of various types of sampling procedures that can be adopted to ensure that individual biases or the influences that you can have because of your personal relationship with the samples can be addressed to a major extent, right. So, you can refer to this in detail and understand the structure.

(Refer Slide Time: 18:17)



But now move on to the sample size. Now, this is one of the most important topics, whenever I talk about usability; my students and whenever I give lectures to any other places, people ask about what is the ideal size of samples that are required for a usability testing experiment. Now, if you go back to literatures, if you go back to what literatures who have done extensive scientific research on samples and the sizes; you would understand that they talk about a minimum guideline for specific research methods.

Like for conducting interviews, if you have more than 12 participants, that is an ideal structure; but often it is said in in many of my studies that I have conducted, I have seen that after 6, 7 participants if the participants are providing, you really good data, you can you can see occurrence recurrence or specific themes. And by the time you go reach 12, you really have a rich data that is more repetitive in nature and you can understand a specific structure from that data.

Now, if we are conducting one tailed experiment. Now, what is a one-tailed experiment? For us to understand one tailed experiment, we must go back to what we have talked about in hypothesis. As we have discussed about hypothesis, which is a prediction or an assumption and in hypothesis what we do; we say that if we use these interface feature or this interface feature say 1, would increase satisfaction scores.

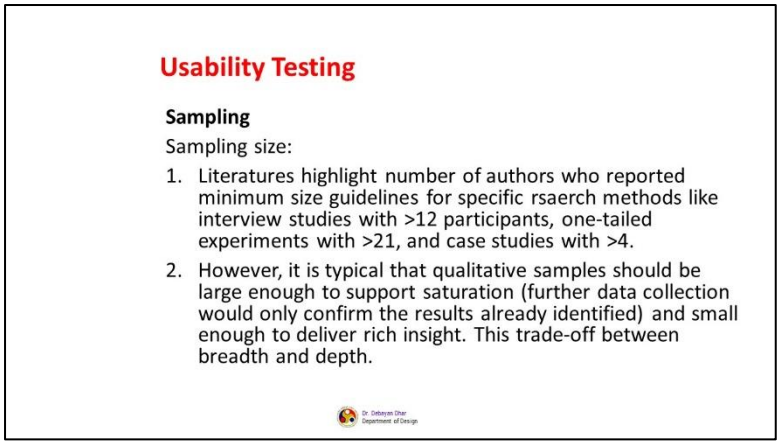
Or we can say it would reduce errors right, that is a hypothesis. We want to see that in this interface we have incorporated a new feature and because of this feature, we are predicting that the errors will get reduced; the satisfaction course will improve, you know that is our prediction, that is our assumption right and we are going to taste it. Now, this is called a hypothesis which is one tailed. One tailed means what?

Now, to understand one tailed, you need to understand how data is distributed. When we collect data from samples ideally, it should get distributed like this and this is called a bell-shaped curve or a normal distribution, right. Now, here what you see is, this is the mean value and these are the extreme data's in either side of the mean value. Now, in this situation, if we say that the interface features 1 would increase the satisfactions course; we are only talking about a particular side of data.

And we are predicting that because of that feature, there is always this chance of that this mean can shift towards this, because of increase in this type of data, which is positive in nature; that means positive satisfactions course. Now, this is one tail. If you see this normal distribution, this side is considered to be one tail and this is the other side that is considered to be the other tail.

Now, if we had said that because of this feature, this interface feature might influence the satisfaction; that means we are not sure whether the influence would be a positive influence or a negative influence, that means we are talking of both the sides, right.
The positive side of it or the negative side of it; it can influence both the side, then this kind of hypothesis is called a two tailed hypothesis. But if we are referring to if our hypothesis is that; we are pretty sure that, it is not going to reduce the satisfaction course, but we are predicting that that it will increase satisfactions, that means we are only talking about this side. So, if this data increases, this mean will shift; this is the existing data and we are we would be comparing this data with another data, which is collected because of a new interface.

(Refer Slide Time: 23:58)



And this new interface what we would see that, if the mean value lies at 4 here, it would shift to 6 here; the positive change that has happened in terms of increase in satisfactions course. Now, this is called a one tailed test. So, in cases where we are focusing on one-tailed experiments, which only predict one side; either it will increase or decrease, we are not situations where we are not sure whether there would be increase or decrease.

We are sure that there would be a difference, but we are not sure about which side the difference would happen; those are two- tailed experiments. But if we are sure about that it might increase or it might decrease, then it is called a one tailed experiment. Now, in cases of one-tailed experiments, the ideal number many literatures say is that, anything that is more than 21 will work.

Now, do not take this as a matter of fact that, every time if you collect anything beyond 21 would be working. In some cases, if your samples do not provide accurate data, in those cases you need to see whether the distribution is coming to be normal or not. If the distribution is not coming normal, then you have to might increase the number of data.

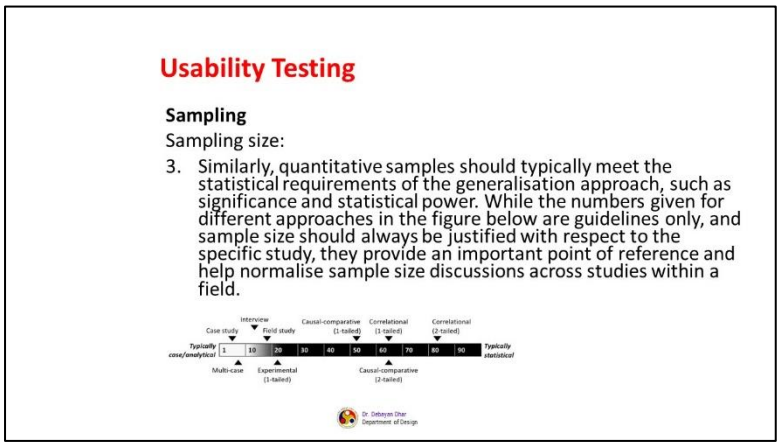
And as you collect more data, you would see your distribution would tend towards a normal distribution. Any case studies if you are working, anything more than 4 ideally; but these are

some of the benchmarks that have been used across literatures. But the more number of samples if you collect, the more quality of data you have and the more powerful is your prediction about the population parameter.

Now, in qualitative samples, should have a large enough data to support the saturation. What we mean by saturation? We mean by saturation that, the themes that are being generated from the qualitative data; as you start getting more and more samples, you would see that no new themes can be generated, same themes are being recurring, that is called saturation.

Further data collection would only confirm the results already identified and small enough to deliver reach in sight; this is the trade-off that you have to do between breadth and depth of data.

(Refer Slide Time: 26:50)



Now, quantitative samples should typically meet the statistical requirements of generalization approach, such as significance and statistical power. While the numbers given for different approaches in the figure below that you see here, which can act as a reference are guidelines only and sample size should always be justified with respect to the specific study; they provide an important point of reference and help normalize samples as discussions across studies within

the field.

Now, what is meant here is that consider the situation that; I often give this example in my class that, you have gone to the market and you want to buy some vegetables and there is a stack of vegetables and you start pulling out each one of the vegetables one by one to see its quality, right.

Now, it might happen say that in that bag, in that basket, there are three bad vegetables; let us say apples, three bad apples. Now, incidentally you pulled up those three apples only which are bad. So, one by one as you start pulling them out and you see that the apples are have rotten, a rotten apple or they are bad; you tend to believe that all these apples of this basket will be bad, is not it if it happens thrice.

But in real case what might happen that, these were only the three apples which you have taken from the basket, which were incidentally bad, rest of the apples were good. But then your activity of pulling three apples out from the basket and arriving to a conclusion that, the entire apples are bad is a wrong conclusion.

So, in research also, since we are extracting samples from the population; it might happen that the samples may have issues in themselves. There can be individual issues characteristics, there can be other environmental characteristics, there can be many things; that because of which your samples have not provided accurate data or the data which you have gathered cannot be relied upon.

Now, in such scenarios, you need to identify a possible case; that this is the case if, this is the situation I am going to believe, if the situation exceeds this value, then I am not going to believe this data. And what is that structure?

That is what we call statistical significance. So, we say out of this 100 samples, if 5 of them is poor or anything below 5; anything below 5 samples if they are not correct, if they are biased, if they are rotten, that is ok for me. But anything more than 5 is not something that we are going to accept; because then, then there is an issue with the samples, that is called this statistical significance.

So, here you have talked about what 95 percent, 5 percent you said; when you say 5 out of 100, you are talking about 5 percent of the data, which you are ready to except that. If it is within 5 percent range, I am going to accept that the samples are ok. But if it is beyond that range, then I am going to say that they are not, ok. When you compare two groups; see in split half testing, we talk about comparison of two groups and we have normal distribution across these groups.

Say this is the normal distribution that you have group A and group B. What we are supposed to do is that? We are supposed to compare between these means and if we say that more than 5 percent of the solutions does not come; more than 5 percent of the solutions of the data in group B, that are not much difference among those data, then we consider that we did not have identified a significant difference across the two groups, ok.

That also tells us that the samples are distributed in such a way that a significant difference between group A and group B cannot be observed and that is what we call significance, statistical significance.

So, ideally in situations say for example, this number varies across study subjects and study situations; if it is something where you are testing a drug, say vaccines, then even one sample if it does not show effective changes, then it might be of concern. But in our cases, we can say that ok 5 percent, 10 percent if they show; if they do not show any changes its ok, if 90 percent of them are showing different means, right.

Then we are going to consider that our experiment is being successful and we can extrapolate that the population parameter is going to get effected, because of the changes that we have done in our interface designs. And here apart from statistical significance, we also talk about this statistical power of the test, right.

The power means the ability of your test; power when we talk about statistical power, we mean that the ability of your experiment to have significant differences across the samples in two groups, that are in question and these becomes the bed rock for your studies.

So, one is how much difference both groups has and whether the difference are sufficient enough for you to predict that the interface design that you have conceived has significant, has been conceived as a significant difference and because of which the metrices and the results from the experiments are different.

And when we talk about the statistical power, we are eventually talking about the ability of your experiment to reject that or to accept the fact that there is a difference between the two groups and we are rejecting the fact that there is no difference between the two groups. And when we say that our assumption, we are rejecting our assumption that there is no difference between the two groups and that means, that we are talking about assumption that there is a difference that exists between the two group's results.

So, coming back to what we define the assumptions as and we have talked about hypothesis, generally hypothesis also have two different types; one is the null hypothesis and the other one is the alternate hypothesis. So, what we have talked about that, these interface feature effects will increase the satisfaction score, this is an alternate hypothesis.

Specifically, the dull hypothesis is a statement that says that there is no difference or there is no effect; that means even if you say that if we have a new design, it is not going to significantly make it different across the data or the satisfactions course. Say for example, you have concept A and concept B and somehow you have a hunch that the concept B or the concept; one of the concepts say concept B is much more superior in terms of certain values and therefore, it is going to be highly rated.

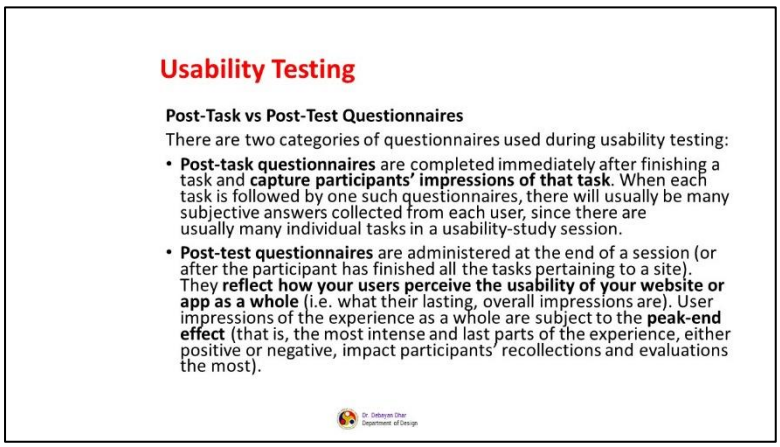
But to prove that you first take a stand that ok, I first of all let me see that these two concepts and the if I give it to my samples and if they are going to work on them; then there is no difference in terms of errors, in terms of time on completion, in terms of satisfactions course that we are getting, that is the statement you are going to reject.

And at any point of time when you see there a difference that exists, you focus on the significance of that difference; whether it comes in the boundary that you have defined or not and then you say that ok, I am now satisfied with the data and I can clearly see that there is a

difference and that our hypothesis which we call as the null hypothesis that there is no difference is not, is incorrect.

And we are going to reject this null hypothesis and we are going to say now that there is a difference. And this difference is significant in nature; because majority of the data exhibits something that is 95 percent or more than 95 percent, that is what statistical significance means, right.

(Refer Slide Time: 37:27)



Now, let us come down to the concept of pre-task, post task and post-test questionnaires. Now, as we have discussed about that questionnaires are the instruments which you are going to use for extracting data. And these data are going to tell you, whether your users are satisfied with your interface or not, there can be other course, for example task load questionnaires also; not only satisfactions course can be extracted by questionnaires, but can be task load can also be extracted from questionnaires.

Now, the two categories are post task questionnaires and post-test questionnaires. Now, post task questionnaires are completed immediately after finishing a task. So, immediately a task is

given. So, you have given a task to your samples and after this task you provide them with a question and they would fill up the question and give you the data and they would finish the questionnaire filling; that is called post task questionnaires, it captures participants impression of that task.

So, when each task is followed by one such questionnaires, they will usually be many subjective answers collected from each user. And because there would be many individuals in the samples in the usability studies session and that is called a post task questionnaire, it is focused on a particular task.

What is the post-test questionnaire? Now, post-test questionnaires are administered at the end of the entire session; it is now a session can be comprising, a session can comprise of multiple tasks. And multiple tasks together can comprise is actually one session, it might be; not necessarily that every time a session comprises of multiple tasks, it can be one task also.

Now, when you administer, you know after the participant has finished all the tasks pursuing to the a concept on the side of the software; when you administer a questionnaire at the end of the entire session, it is called a post-test questionnaire. And the reflect how your users perceive usability of your sight or app as a whole.

Now, your software can have multiple tasks and after each task, you collect; you give them a questionnaire, they fill up the questionnaire based on the impressions of the task. And after they have completed the all the major task that your software does, after the end of the entire session; you give the post-test questionnaire that, would reflect how your users perceive the usability of your website or the application as a whole. Now, what their lasting and overall impressions are that is what is valuable to you.

So, users impression of the experience as a whole are subject to the peak end rule that we have talked about. If you remember in our earlier lectures, we have talked about this peak end rule and which is that the most intense and last parts of the experience; it can be either positive or negative, impact participants recollections and evaluations the most.

We talked about the pain related experiments earlier and that is what we are focusing here as

well. So, that means we are more concerned about the experience that our samples are getting at the fag and at the last part of the interaction with the entire system; it can be positive and negative and this is going to shape their overall experience and therefore, post-test questionnaires are so important.