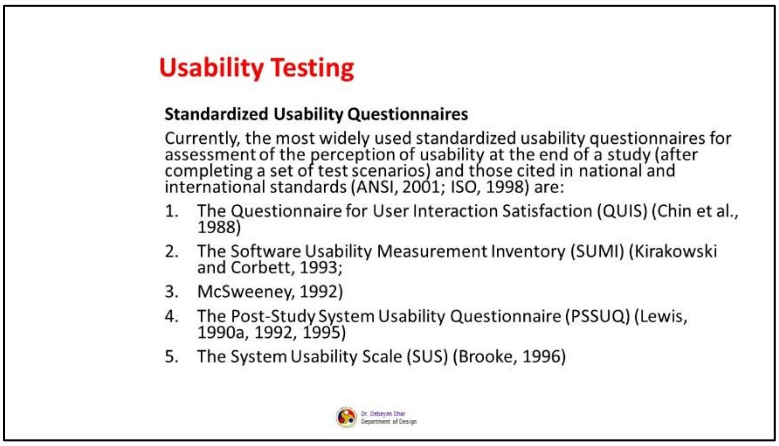
Usability Engineering
Dr. Debayan Dhar
Department of Design
Indian Institute of Technology, Guwahati

Module - 12
Lecture - 36
Usability Testing

(Refer Slide Time: 00:32)



Usability Testing

Standardized Usability Questionnaires

Currently, the most widely used standardized usability questionnaires for assessment of the perception of usability at the end of a study (after completing a set of test scenarios) and those cited in national and international standards (ANSI, 2001; ISO, 1998) are:

1. The Questionnaire for User Interaction Satisfaction (QUIS) (Chin et al., 1988)
2. The Software Usability Measurement Inventory (SUMI) (Kirakowski and Corbett, 1993;
3. McSweeney, 1992)
4. The Post-Study System Usability Questionnaire (PSSUQ) (Lewis, 1990a, 1992, 1995)
5. The System Usability Scale (SUS) (Brooke, 1996)

Dr. Debayan Dhar
Department of Design

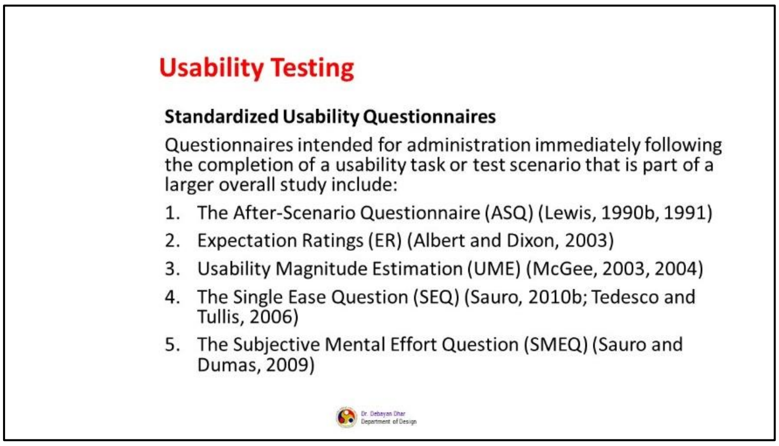
So, we will continue discussion on Usability Testing. And we will be discussing about various questionnaires that are being used as part of a measure to capture the data you know after the entire task completes or the entire session completes. So, what are the standardized questionnaires that are used or referred by user experience design practitioners or human computer interaction practitioners we will discuss about that.

Now, some of the most widely used and standardized questionnaires I have listed for all of you here in this slide. You can see the first one is the questionnaire for user interaction satisfaction that is called QUIS, then the software usability measurement inventory SUMI, then we have the post study system usability questionnaire which is called as PSSUQ and the system usability scale.

So, out of this 4 this one if you see this is the 1 2 this is wrongly here given the number this is the third and the fourth one ok. Now, out of these 4 scales this one is the most widely used one though all of them are widely used one. But this if you go to research papers and read research papers you would see a lot of researchers in this industry and in this field, they use The System Usability Scale by Brooke of in which was published in 1996.

Now, these standardized usability questionnaires are used for assessment of the perception of usability at the end of the study. So, these are all about perception how do your user perceive usability of a particular product, about a particular activity that your concept is providing with the users and these are some of the standardized one.

(Refer Slide Time: 02:42)



Usability Testing

Standardized Usability Questionnaires

Questionnaires intended for administration immediately following the completion of a usability task or test scenario that is part of a larger overall study include:

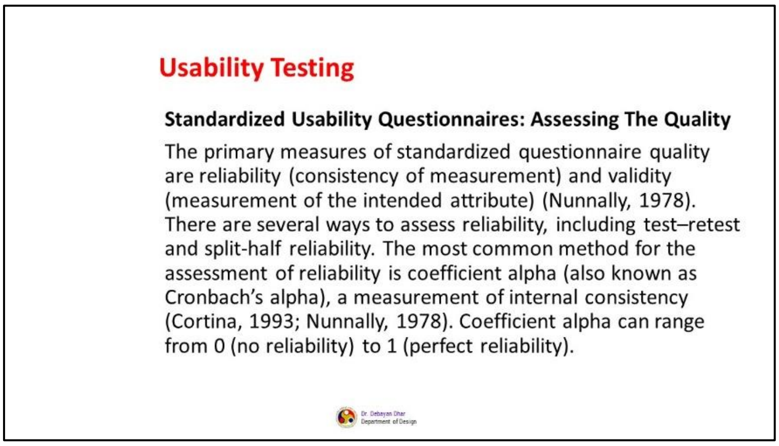
1. The After-Scenario Questionnaire (ASQ) (Lewis, 1990b, 1991)
2. Expectation Ratings (ER) (Albert and Dixon, 2003)
3. Usability Magnitude Estimation (UME) (McGee, 2003, 2004)
4. The Single Ease Question (SEQ) (Sauro, 2010b; Tedesco and Tullis, 2006)
5. The Subjective Mental Effort Question (SMEQ) (Sauro and Dumas, 2009)

Dr. Debayan Dhar
Department of Design

Now, there are questionnaires which are intended for administration immediately following the completion of a usability task also.

Some of the scales that we have we talked about after scenario setting and also the entire case scenario. Some of the are some of them are the after-scenario questionnaire by Lewis, expectation ratings usability magnitude estimation the single use question scenario questionnaire and a subjective mental effort questionnaire. We will discuss about some of them in detail.

(Refer Slide Time: 03:20)



Usability Testing

Standardized Usability Questionnaires: Assessing The Quality

The primary measures of standardized questionnaire quality are reliability (consistency of measurement) and validity (measurement of the intended attribute) (Nunnally, 1978). There are several ways to assess reliability, including test–retest and split-half reliability. The most common method for the assessment of reliability is coefficient alpha (also known as Cronbach's alpha), a measurement of internal consistency (Cortina, 1993; Nunnally, 1978). Coefficient alpha can range from 0 (no reliability) to 1 (perfect reliability).

Dr. Debayan Dhar
Department of Design

Now, why are we harping on this thing the word which is being used here standardized? Why we are talking about standardized usability questionnaire here? Now, understand that the primary measures of standardized questionnaire quality are what reliability; that means consistency of measurement. Every time you are measuring it should measure the same construct that for which it is intended to measure. And the next important parameter is validity, measurement of the intended attribute or the construct.

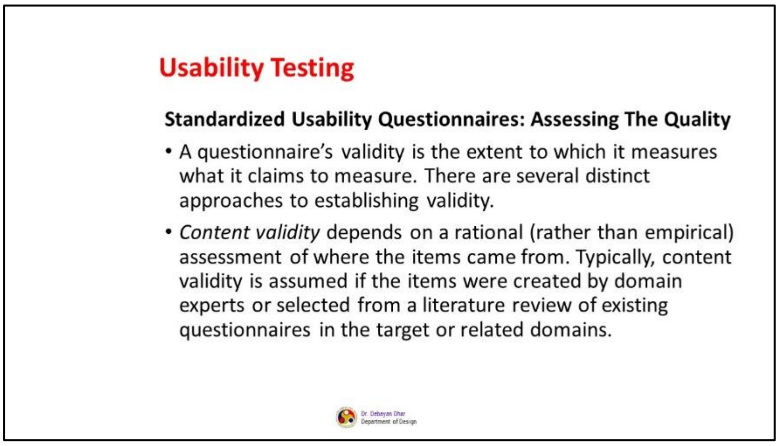
Now, there are several ways to assess reliability some of them include test retest and split half reliability. But the most common method for the assessment of reliability is coefficient alpha which is also called as Cronbach's alpha and which is a measurement of internal consistency of your questionnaire or the instrument. Now, coefficient alpha can range from 0 to 1, 0 means there is no reliability and 1 means perfect reliability.

So, a good standard at a questionnaire generally has the range from 0.75 to 0.8 or 0.7 that is an ideal range in which your Cronbach's alpha would rely on. Now, one of the important aspect which you need to understand here is that we are talking about quantitative parameters all these instruments, all these questionnaires that are being discussed these are questionnaires that capture data as a quantitative value.

We are not capturing data as qualitative parameter and therefore, we can then measure the what we talk about the internal consistency which is the reliability of your instrument. Now, if your instrument is not reliable what will happen is that every time you go and conduct a test and you use these questionnaires it would not provide you with the same measure again and again.

So, if you are if the questionnaire is about measurement of perceived usability at certain at a certain point of time it can measure the perceived usability while in a different test or a say situation it might measure something else. So, then if it does not measure over the period of time the same internal attribute then there is an issue in consistency of its measurement and that is what we call as reliability issue and this is measured by Cronbach's alpha.

(Refer Slide Time: 06:08)



Usability Testing

Standardized Usability Questionnaires: Assessing The Quality

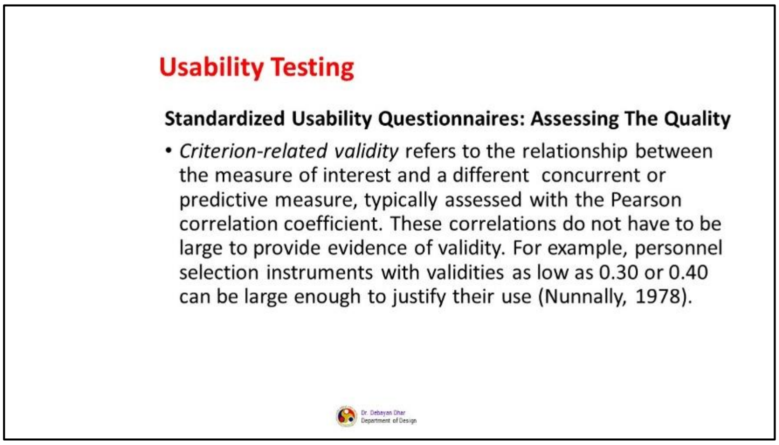
- A questionnaire's validity is the extent to which it measures what it claims to measure. There are several distinct approaches to establishing validity.
- *Content validity* depends on a rational (rather than empirical) assessment of where the items came from. Typically, content validity is assumed if the items were created by domain experts or selected from a literature review of existing questionnaires in the target or related domains.

Dr. Debayan Dhar
Department of Design

Now, a questionnaires validity now coming down to validity it is the extent to which it measures what it claims to measure. That means, if your if you have a questionnaire that measures the perceived usability the questionnaire must measure perceived usability. If you if you have designed a questionnaire that captures or measures you know memorability, if it captures preferences or something like that.

It must capture that construct for which it is being designed and that is what we called as validity, the extent to which it measures what it claims to measure right. Now, there are several distinct approaches to measure validity of a instrument or a questionnaire, the first one is the content validity. Now, content validity depends on a rational it is a subjective evaluation and therefore, it is not empirical in nature. It is a rational assessment of where the items came from. Now, typically content validity is assumed if the items were created by domain experts. So, somebody who has worked in the area of usability they know the measures, the constructs, the parameters or the kind of questions that can be asked to measure usability or these questions are selected from a literature review of existing questionnaires in the target or the related domains. If that is the case then we say that the content validity has been achieved or addressed.

(Refer Slide Time: 07:51)



Usability Testing

Standardized Usability Questionnaires: Assessing The Quality

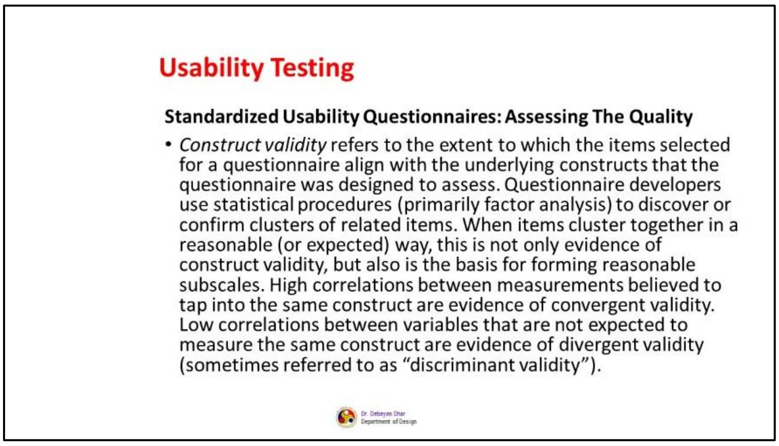
- *Criterion-related validity* refers to the relationship between the measure of interest and a different concurrent or predictive measure, typically assessed with the Pearson correlation coefficient. These correlations do not have to be large to provide evidence of validity. For example, personnel selection instruments with validities as low as 0.30 or 0.40 can be large enough to justify their use (Nunnally, 1978).

Dr. Debayan Dhar
Department of Design

The next one is the criteria related validity. Now, criteria related validity refers to the relationship between the measure of the interest. So, this is the parameter and a different concurrent or predictive measure you know. So, it is a comparison or an assessment between these two parameters that has been conducted.

So, typically it is assessed with the Pearson correlation coefficient. Now, these correlations do not have to be large to provide evidence of validity. Now, for example, say personal selection instruments with validities as low as 0.3 or 0.4 can be large enough to justify their use right, this is called criteria related validity.

(Refer Slide Time: 08:41)



Usability Testing

Standardized Usability Questionnaires: Assessing The Quality

- *Construct validity* refers to the extent to which the items selected for a questionnaire align with the underlying constructs that the questionnaire was designed to assess. Questionnaire developers use statistical procedures (primarily factor analysis) to discover or confirm clusters of related items. When items cluster together in a reasonable (or expected) way, this is not only evidence of construct validity, but also is the basis for forming reasonable subscales. High correlations between measurements believed to tap into the same construct are evidence of convergent validity. Low correlations between variables that are not expected to measure the same construct are evidence of divergent validity (sometimes referred to as "discriminant validity").

Dr. Debayan Dhar
Department of Design

The third one is construct validity. Now, construct validity refers to the extent to which the items selected for a questionnaire align with underline constructs that the questionnaire was designed to assess.

So, questionnaire developers use statistical procedures like primarily factor analysis to discover or confirm these clusters of the related items right. So, when items cluster together in a reasonable or an expected way this is not only evidence of construct validity, but also is the basis of forming reasonable sub scales which we will talk about in detail in subsequent slides.

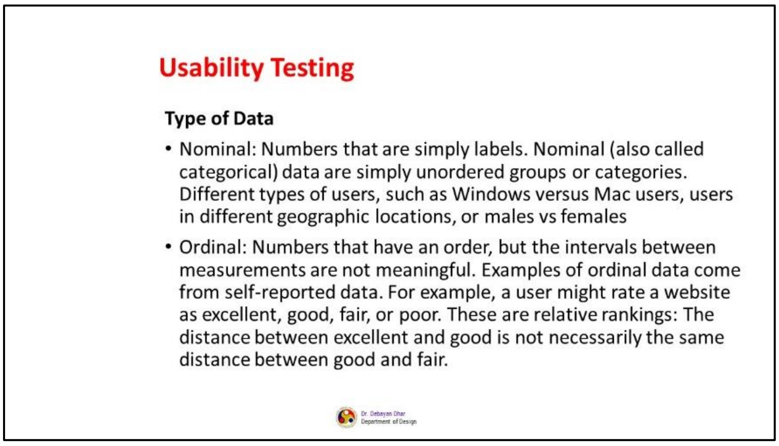
Now, high correlations between measurements believed to tap into the same construct are evidence of convergent validity. That means, if some of the sub parameters say P 1, P 2 if some of these sub parameters are being seen as correlated. Then we can conclude by saying that the same there is an evidence for convergent validity means they are converging towards a particular construct right.

So, low correlations between variables that are not expected to measure the same construct are evidence of divergent validity. So, both a way the validities can be measured. So, construct validity can be measured both ways using convergent validity like the sub parameters are directly are getting correlated with the other parameters or we are using an opposite construct and looking at whether these constructs are or the study suggests whether there is any divergent correlation.

That means they are not at all correlated and there is a divergence in terms of the parameters that are under investigation and that is an evidence of divergent validity. So, low correlations sometimes this is often referred to as discriminant validity as well. So, measurement of validity what we have understood is can be carried out in three ways one is the content validity that is primarily rational in nature it is not empirical or quantitative.

The second one is criteria related validity which can be assessed by the Pearson's coefficient. Then third one is a construct validity which is based on how other sub parameters tend to correlate with the major parameter in terms of their correlation activity. So, therefore, if there is a high correlation we talk it we say that there is an evidence of convergent validity.

If there is low correlation then we of different construct of constructs that we know that are not related then it is an evidence of divergent validity which is also called as discriminant validity. (Refer Slide Time: 11:55)



Usability Testing

Type of Data

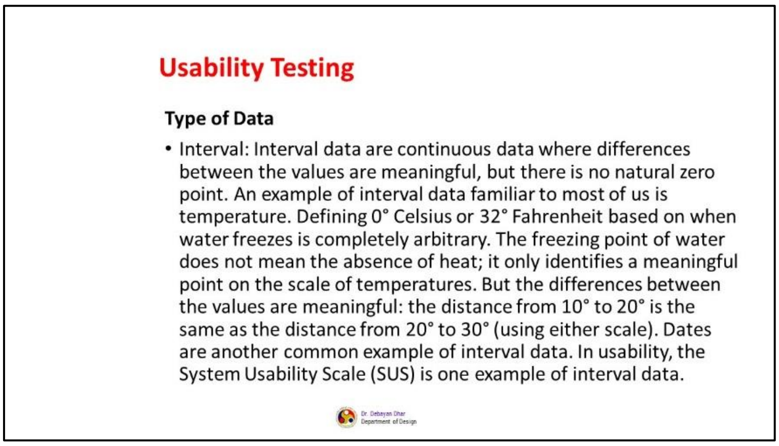
- Nominal: Numbers that are simply labels. Nominal (also called categorical) data are simply unordered groups or categories. Different types of users, such as Windows versus Mac users, users in different geographic locations, or males vs females
- Ordinal: Numbers that have an order, but the intervals between measurements are not meaningful. Examples of ordinal data come from self-reported data. For example, a user might rate a website as excellent, good, fair, or poor. These are relative rankings: The distance between excellent and good is not necessarily the same distance between good and fair.

Dr. Debayan Dhar
Department of Design

Now, coming down to finally, covering all the major aspects of your research design of how do you conceive a usability test scenario one of the fundamental parameters that are of concern and that would let you know the kind of tests that you are going to do is based on the type of data.

Now, when we talk about type of data we mean what? We mean that the scales that you are using or that the construct that you are measuring it has a particular set of value and that value can be classified or categorized into four different types.

(Refer Slide Time: 12:41)



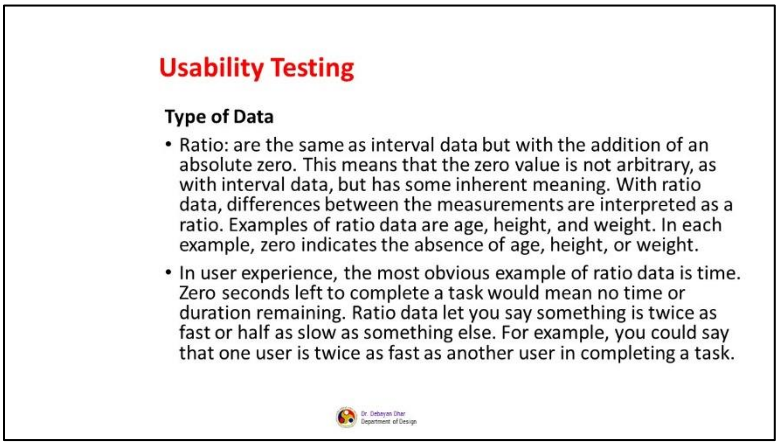
Usability Testing

Type of Data

- Interval: Interval data are continuous data where differences between the values are meaningful, but there is no natural zero point. An example of interval data familiar to most of us is temperature. Defining 0° Celsius or 32° Fahrenheit based on when water freezes is completely arbitrary. The freezing point of water does not mean the absence of heat; it only identifies a meaningful point on the scale of temperatures. But the differences between the values are meaningful: the distance from 10° to 20° is the same as the distance from 20° to 30° (using either scale). Dates are another common example of interval data. In usability, the System Usability Scale (SUS) is one example of interval data.

Dr. Debayan Dhar
Department of Design

(Refer Slide Time: 12:44)



Usability Testing

Type of Data

- Ratio: are the same as interval data but with the addition of an absolute zero. This means that the zero value is not arbitrary, as with interval data, but has some inherent meaning. With ratio data, differences between the measurements are interpreted as a ratio. Examples of ratio data are age, height, and weight. In each example, zero indicates the absence of age, height, or weight.
- In user experience, the most obvious example of ratio data is time. Zero seconds left to complete a task would mean no time or duration remaining. Ratio data let you say something is twice as fast or half as slow as something else. For example, you could say that one user is twice as fast as another user in completing a task.

Dr. Debayan Dhar
Department of Design

And these are nominal data, ordinal data interval, data and ratio data majority of the time interval and ratio can be clubbed together and be called as a scale data.

Now, what is the nominal data? So, numbers that are simply levels for example, you have a jersey number or you know that are also called categorical in nature. So, nominal data is also called categorical in nature and these are simply unordered groups or categories different types of users such as Windows versus Mac users, users in different geographic locations or males versus females, these are examples of nominal data.

Then you have ordinal data. Now, numbers that have an order for example, your roll number, but the intervals between the measurements are not meaningful. So, when we say that what was the interval between roll number 1 and roll number 2 we cannot ascertain that the same interval exist between all these roll numbers right. So, examples ordinal data come from self-reported data.

For example, user might rate a website as excellent, good, fair, or poor. Now, we have no idea the intervals between excellent and good or good and fair or fair or poor. Now, these are relative rankings the distance between excellent and good is not necessarily the same distance between good and fair. And therefore, these are called ordinal data, but they can be ranked. So, if you rank it can be excellent, good, fair, or poor.

So, though it can be ranked the intervals between them are not consistent in nature, but in case of nominal data it cannot be ranked as well you cannot rank between gender males or females right. The next one is called the interval data. Now, in interval data are continuous data that where differences between the values are meaningful in nature. Now, if you remember the last one we talked about categorical data and continuous data.

Continuous determines what? The data where these differences these intervals are meaningful; that means, the interval between 1 to 2 is same as 2 to 3 or 3 to 4 right and therefore, these are called as continuous data, but there is no natural 0 point. So, an example of interval data familiar to most of us is the concept of temperature.

Now, when we talk about say temperature like 0 degree Celsius or 32-degree Fahrenheit which is based on what water freezes is completely arbitrary it is a it has a frame of reference right. So, the freezing point of water does not mean the absence of heat it only identifies a meaningful point on the scale of temperatures. But the differences between the values are meaningful the distance from 10 degree to 20 degree is the same as a distance between from 20 degree to 30 degree using either of the scales.

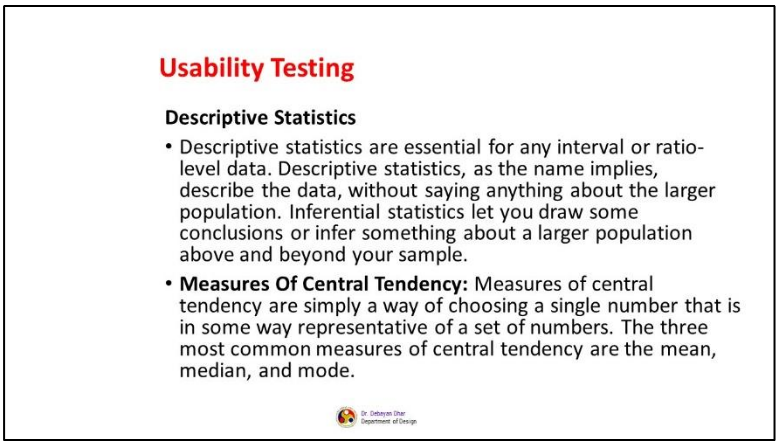
Either you use the Celsius scales or the Fahrenheit scales. Now, dates are another common example of interval data and in usability the system usability scale which we talked about has one of the most highly used scales, most popular among all these scales that are being used by user experience designers and HCI designers is an example of an interval data.

The last one is ratio data now, what is the ratio data? So, ratio data are the same as interval data, but the addition of an absolute 0. So, the concept of absolute 0 exists it is not in terms of the frame of reference that we had in case of an interval data. This means that 0 is value is not arbitrary in nature as with the interval data, but has some inherent meaning. So, with ratio data differences between the measurements are interpreted as ratio.

Examples ratio data are age, height, weight, right in each example 0 indicates the absence of age or the absence of height or the absence of weight. Now, these are the classifications of the different types of data that would let us understand what kind of measurement or analysis techniques we need to proceed based on the interpretation of the type of data from our scales.

Now, in user experience the most obvious example of ratio data is time ok. So, 0 seconds left to complete a task would mean no time or duration remaining. So, ratio data let you say something is twice as fast or half as low as something else is in reference to an arbit not an arbitrary data, but an absolute data which we stock about as 0; that means, nothing is left. So, for example, you could say that one user is twice as fast as another user in completing a task.

(Refer Slide Time: 18:00)



Usability Testing

Descriptive Statistics

- Descriptive statistics are essential for any interval or ratio-level data. Descriptive statistics, as the name implies, describe the data, without saying anything about the larger population. Inferential statistics let you draw some conclusions or infer something about a larger population above and beyond your sample.
- **Measures Of Central Tendency:** Measures of central tendency are simply a way of choosing a single number that is in some way representative of a set of numbers. The three most common measures of central tendency are the mean, median, and mode.

Dr. Debayan Dhar
Department of Design

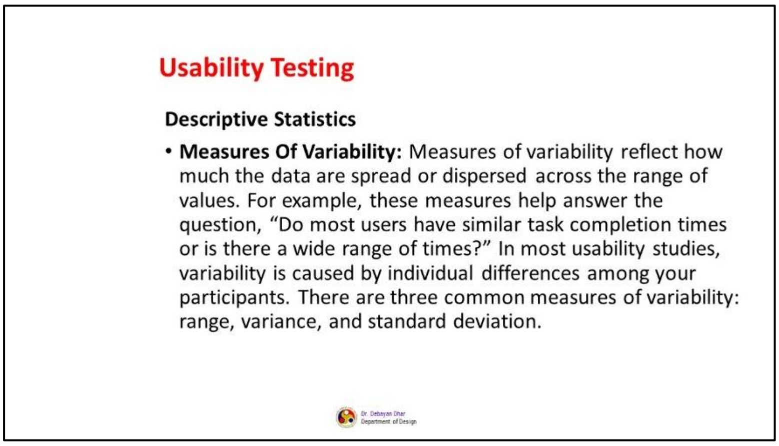
So, having understood all these types of data now what is important is to understand the concepts of the descriptive statistics and how you compare between two different groups. Now, when we talk about descriptive statistics now it is it is essential to understand that for any interval or ratio level data descriptive, reporting descriptive statistics is important. Now, descriptive statistics as a name implies you know it describes the data.

It tells you about the data without saying anything about the larger population. See we have we have collected some samples from the population and we are trying to understand that our samples represent our population. So, whatever we do or test or is being provided as feedback by our samples we will be able to gauge how the situation would exist in the population right. So, inferential statistics let you draw some conclusions or infer something about larger population above and beyond your sample.

So, while descriptive statistics allows us to understand about the nature of the sample using samples to extrapolate the population parameter is what we can we term as the concept of inferential statistics. It allows us to infer about the population parameter. And some of the ways through which we use inferential statistics are measures of central tendency. So, measures of central tendency simply are way of choosing a single number that is in some way representative of a set of numbers.

So, the three main most common references or measures of central tendency are the mean, median and mode. And remember that which measure of central tendency would you consider would depend on the type of data that we have just talked about nominal, ordinal, interval or ratio.

(Refer Slide Time: 20:04)



Usability Testing

Descriptive Statistics

- **Measures Of Variability:** Measures of variability reflect how much the data are spread or dispersed across the range of values. For example, these measures help answer the question, “Do most users have similar task completion times or is there a wide range of times?” In most usability studies, variability is caused by individual differences among your participants. There are three common measures of variability: range, variance, and standard deviation.

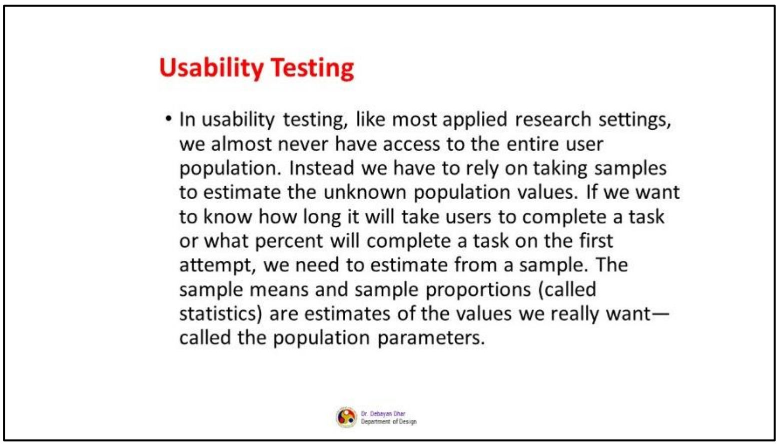
Dr. Debayan Dhar
Department of Design

So, measures of variability, now measures of variability reflect how much the data is spread.

See if you remember we had a discussion earlier that when you collect this kind of interval or ratio data this data ideally should be distributed like this and this is called the normal distribution or the bell-shaped curve right. Now, this is called a distribution and a bell shaped cap. So, the measures of variability reflect how much the data is spread this spread this entire spread from where it started and where it went this entire spread is the is what variability talks about right.

So, for example, these measures help answer the question do most users have similar task completion times or is there a wide range of times if we are measuring task completion times. In most usability studies variability is caused by individual differences; that means, the psychological constructs among your participants. And there are three common measures of variability one is the range, one is the variance and the other one is the standard deviation.

So, the three common measures of variability is range, variance and the standard deviation.
(Refer Slide Time: 21:38)



Usability Testing

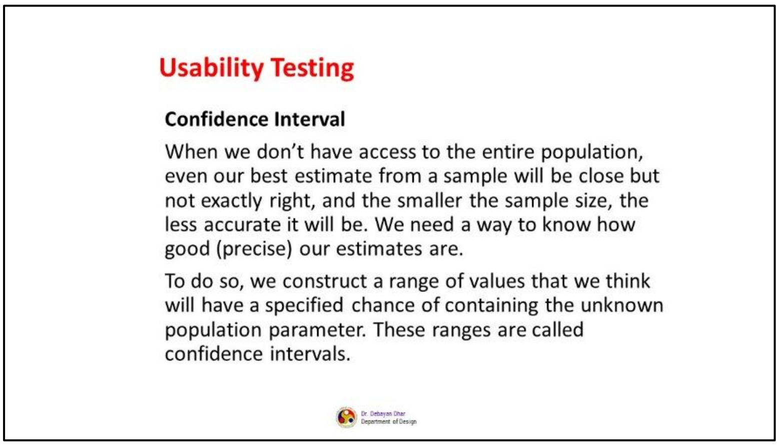
- In usability testing, like most applied research settings, we almost never have access to the entire user population. Instead we have to rely on taking samples to estimate the unknown population values. If we want to know how long it will take users to complete a task or what percent will complete a task on the first attempt, we need to estimate from a sample. The sample means and sample proportions (called statistics) are estimates of the values we really want—called the population parameters.

Dr. Debayan Dhar
Department of Design

Now, in usability testing you know we almost never have access to the entire user population because the user population is really large enough. So, instead what we have to do is we have to rely on taking samples to estimate the unknown population values. So, if we want to know how long it will take users to complete a task or what percent will complete a task on the first attempt we need to estimate from our samples that are representative of our population parameters.

So, the sample means and the sample proportions which are called the statistic are estimates of the values that we really want which we want as something that we call as population parameters.

(Refer Slide Time: 22:24)



Usability Testing

Confidence Interval

When we don't have access to the entire population, even our best estimate from a sample will be close but not exactly right, and the smaller the sample size, the less accurate it will be. We need a way to know how good (precise) our estimates are.

To do so, we construct a range of values that we think will have a specified chance of containing the unknown population parameter. These ranges are called confidence intervals.

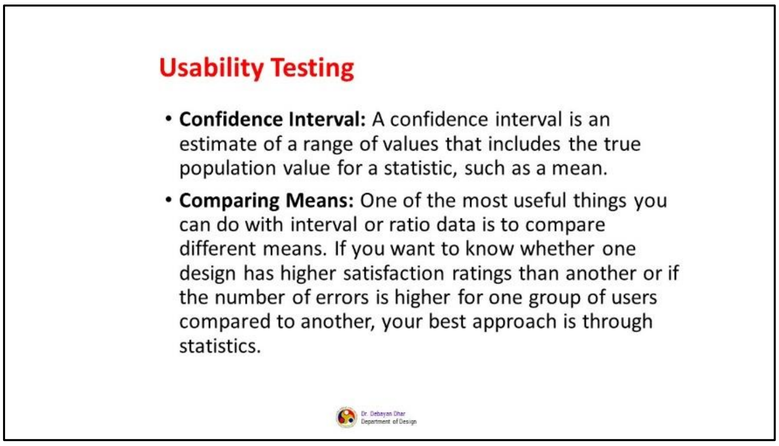
Dr. Debayan Dhar
Department of Design

Now, in this case confidence interval plays a major role.

So, when we do not have access to the entire population even our best estimate from a sample will be close, but not exactly right we cannot pinpoint a number that our mean in the population will lie here, that is absolutely not possible. So, in that case and the smaller sample size the less accurate it will be. So, we need to know how good or precise our estimates are.

So, to do so we construct a range of values that we think will have a specified chance of containing the unknown population parameter and these ranges are called confidence intervals. The bigger your confidence interval is the higher chance that there is a concept of error or an error that has happened in your experimental studies.

(Refer Slide Time: 23:20)



Usability Testing

- **Confidence Interval:** A confidence interval is an estimate of a range of values that includes the true population value for a statistic, such as a mean.
- **Comparing Means:** One of the most useful things you can do with interval or ratio data is to compare different means. If you want to know whether one design has higher satisfaction ratings than another or if the number of errors is higher for one group of users compared to another, your best approach is through statistics.

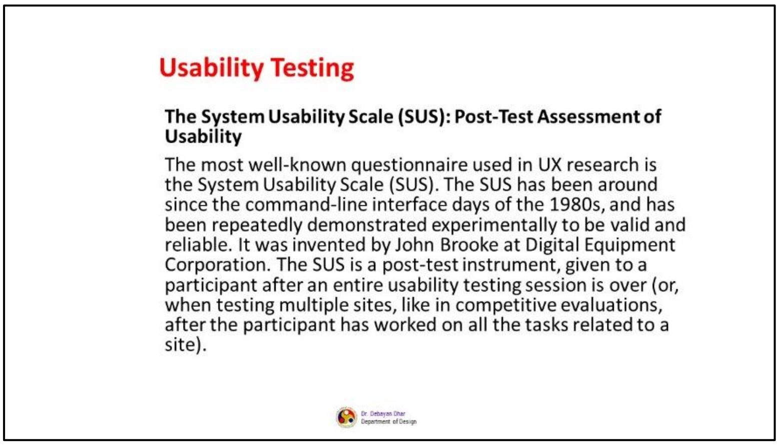
Dr. Debayan Dhar
Department of Design

Now, a confidence interval is an estimate of a range of values that includes the true population value of statistics such as the mean.

So, it means that in this range that is being given from you from your statistical analysis inside that range the true mean of the population parameter will lie that is what we are predicting.

Now, comparing means one of the most useful things you can do with the interval or ratio data is to compare different means if you want to know whether one design has higher satisfaction ratings than other or if the number of errors is higher for one group of users compared to other the best approach is using two groups and measuring the means.

(Refer Slide Time: 24:08)



Usability Testing

The System Usability Scale (SUS): Post-Test Assessment of Usability

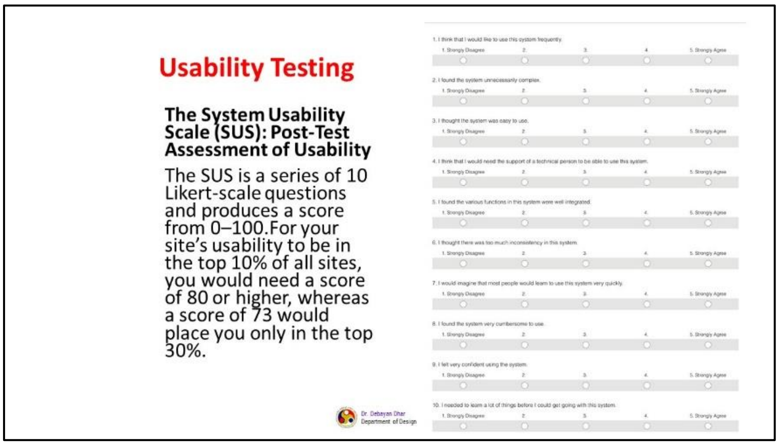
The most well-known questionnaire used in UX research is the System Usability Scale (SUS). The SUS has been around since the command-line interface days of the 1980s, and has been repeatedly demonstrated experimentally to be valid and reliable. It was invented by John Brooke at Digital Equipment Corporation. The SUS is a post-test instrument, given to a participant after an entire usability testing session is over (or, when testing multiple sites, like in competitive evaluations, after the participant has worked on all the tasks related to a site).

Dr. Debayan Dhar
Department of Design

Now, let us come down to the concept these system usability scale that we have just referred to which is used extensively in this domain. Now, the most well-known questionnaire used in UX research or HCI research is the system usability scale.

The system usability scale has been around since 1980s and it was later formally defined by Brooke also and has been repeatedly demonstrated as an experimental to be highly valid and reliable. It was it was invented by John Brooke at Digital Equipment Corporation. And the system usability scale is a post-test instrument which is given to participant after an entire usability testing session is over right.

(Refer Slide Time: 24:57)



Usability Testing

The System Usability Scale (SUS): Post-Test Assessment of Usability

The SUS is a series of 10 Likert-scale questions and produces a score from 0–100.For your site's usability to be in the top 10% of all sites, you would need a score of 80 or higher, whereas a score of 73 would place you only in the top 30%.

1. I think that I would like to use this system frequently.

1. Strongly Disagree	2.	3.	4.	5. Strongly Agree
○	○	○	○	○

2. I found the system unnecessarily complex.

1. Strongly Disagree	2.	3.	4.	5. Strongly Agree
○	○	○	○	○

3. I thought the system was easy to use.

1. Strongly Disagree	2.	3.	4.	5. Strongly Agree
○	○	○	○	○

4. I think that I would need the support of a technical person to be able to use this system.

1. Strongly Disagree	2.	3.	4.	5. Strongly Agree
○	○	○	○	○

5. I found the various functions in this system were well integrated.

1. Strongly Disagree	2.	3.	4.	5. Strongly Agree
○	○	○	○	○

6. I thought there was too much inconsistency in this system.

1. Strongly Disagree	2.	3.	4.	5. Strongly Agree
○	○	○	○	○

7. I would imagine that most people would learn to use this system very quickly.

1. Strongly Disagree	2.	3.	4.	5. Strongly Agree
○	○	○	○	○

8. I found the system very cumbersome to use.

1. Strongly Disagree	2.	3.	4.	5. Strongly Agree
○	○	○	○	○

9. I felt very confident using the system.

1. Strongly Disagree	2.	3.	4.	5. Strongly Agree
○	○	○	○	○

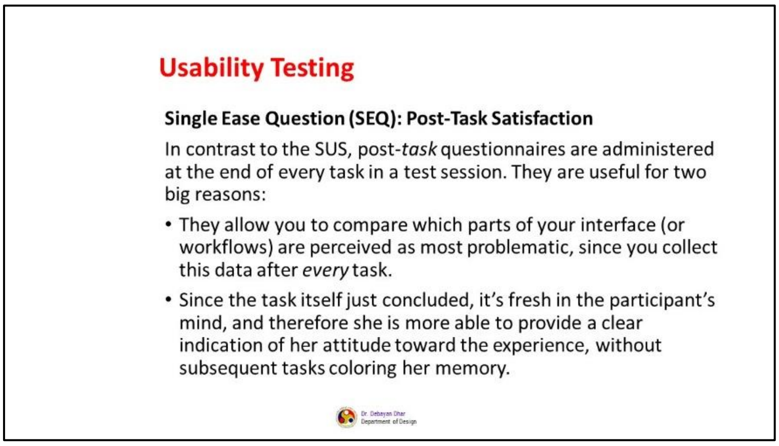
10. I needed to learn a lot of things before I could get going with this system.

1. Strongly Disagree	2.	3.	4.	5. Strongly Agree
○	○	○	○	○

Dr. Debayan Dhar
Department of Design

So, in this slide you can see the example of the system usability scale and the and the questionnaire the test items. The SUS is a series of 10 Likert-scale questions and produces a score from 0 to 100. So, for your sites usability to be in the top 10 percent of all sites you need to have a score of 80 or higher. Whereas, the score of 73 would place you only in the top 30 percent that is the statistic we have from literatures.

(Refer Slide Time: 25:32)



Usability Testing

Single Ease Question (SEQ): Post-Task Satisfaction

In contrast to the SUS, post-*task* questionnaires are administered at the end of every task in a test session. They are useful for two big reasons:

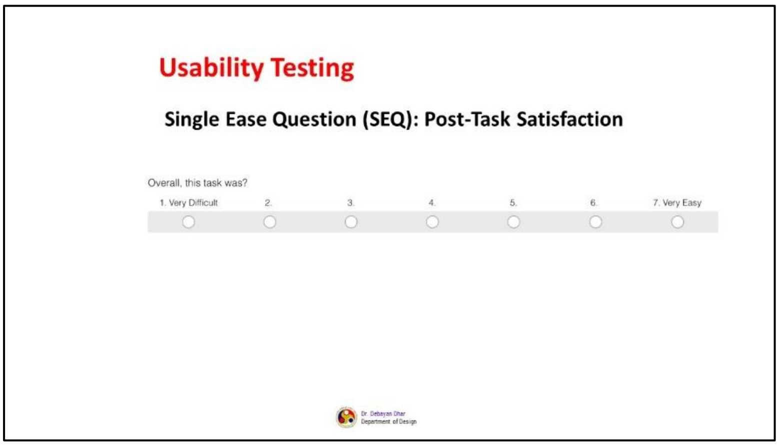
- They allow you to compare which parts of your interface (or workflows) are perceived as most problematic, since you collect this data after *every* task.
- Since the task itself just concluded, it's fresh in the participant's mind, and therefore she is more able to provide a clear indication of her attitude toward the experience, without subsequent tasks coloring her memory.

Dr. Debayan Dhar
Department of Design

Now, we also talked about the single ease question which is a post task satisfaction scale. Now, in contrast to the system usability scale post task questionnaire are administers at the end of every task in a test session. And they are useful for primarily two main reasons and the reasons are that they allow you to compare which parts of your interface are perceived as most problematic since you collect this data after every task right.

So, since the task itself just concluded its fresh in the participants mind. And therefore, she is more able to provide a clear indication of her attitude towards the experience without subsequent tasks coloring her memory.

(Refer Slide Time: 26:16)



Usability Testing

Single Ease Question (SEQ): Post-Task Satisfaction

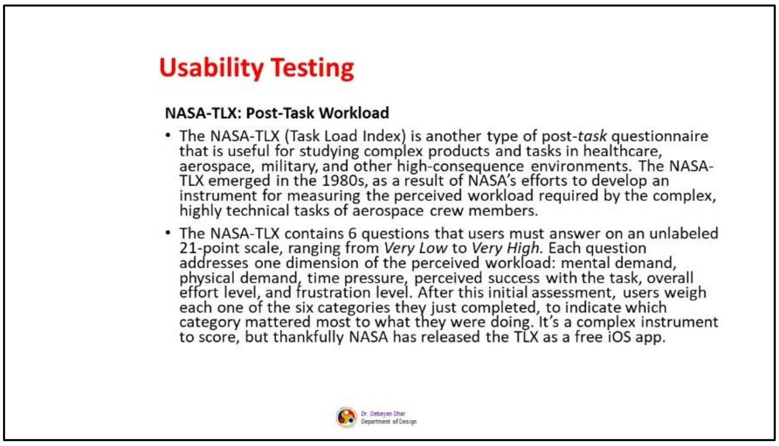
Overall, this task was?

1. Very Difficult	2.	3.	4.	5.	6.	7. Very Easy
○	○	○	○	○	○	○

Dr. Debayan Dhar
Department of Design

Now, this is an example of the single ease question which is post task satisfaction score. So, it is only one item score and it can let you understand about the specific task that has been exposed to the user.

(Refer Slide Time: 26:33)



Usability Testing

NASA-TLX: Post-Task Workload

- The NASA-TLX (Task Load Index) is another type of post-*task* questionnaire that is useful for studying complex products and tasks in healthcare, aerospace, military, and other high-consequence environments. The NASA-TLX emerged in the 1980s, as a result of NASA's efforts to develop an instrument for measuring the perceived workload required by the complex, highly technical tasks of aerospace crew members.
- The NASA-TLX contains 6 questions that users must answer on an unlabeled 21-point scale, ranging from *Very Low* to *Very High*. Each question addresses one dimension of the perceived workload: mental demand, physical demand, time pressure, perceived success with the task, overall effort level, and frustration level. After this initial assessment, users weigh each one of the six categories they just completed, to indicate which category mattered most to what they were doing. It's a complex instrument to score, but thankfully NASA has released the TLX as a free iOS app.

Dr. Debayan Dhar
Department of Design

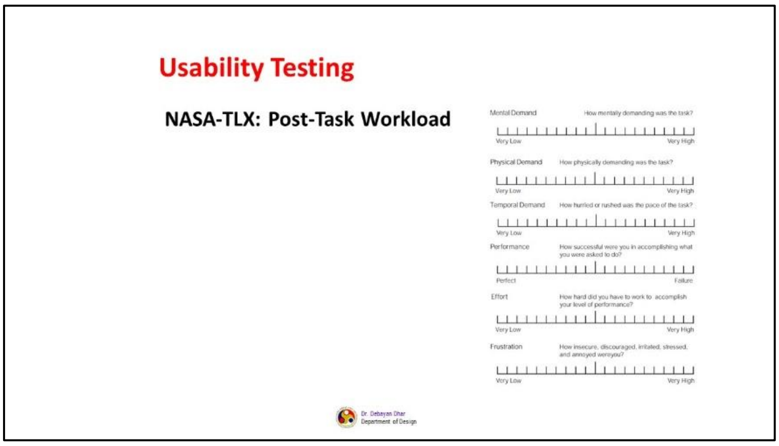
The other questionnaire that we would discuss is a little bit in detail is the NASA task load index. Now, the NASA task load in index is an uh scale that is used to capture the post task works load of the users. So, the NASA TLX which is abbreviated as TLX for task load index is an example of post task questionnaire that is useful for studying complex products and tasks in healthcare, aerospace, military and in and high consequence environments.

So, the it emerged in 1980 from as a result of NASAs efforts to develop an instrument or for measuring the perceived workload working load on the working memory required, but the complex highly technical task of aerospace crew members. But this can be used as extensively in HCI and UX to measure task load of the given tasks that that we have designed and to see how easy it is for our user to remember certain information while they are progressing towards their goal.

Now, the NASA task load index contains 6 questions that users must answer on an unlabelled 21 point scale ranging from very low to very high. Each question addresses one dimension of the perceived workload like mental demand, physical demand, time pressure, perceived success with the task, overall effort level and frustration level. After this initial assessment users weigh each one of the 6 categories they just completed to indicate which category mattered most to what they were doing.

So, it is a complete complex instrument to score, but NASA has released a free online scoring of the task load index as well. It has a free iOS application it is also available in some online interface where you can use the ratings to calculate the final task load and that makes the calculation procedure a little bit easier.

(Refer Slide Time: 28:43)



Usability Testing

NASA-TLX: Post-Task Workload

Mental Demand How mentally demanding was the task?

Very Low Very High

Physical Demand How physically demanding was the task?

Very Low Very High

Temporal Demand How hurried or rushed was the pace of the task?

Very Low Very High

Performance How successful were you in accomplishing what you were asked to do?

Perfect Failure

Effort How hard did you have to work to accomplish your level of performance?

Very Low Very High

Frustration How insecure, discouraged, irritated, stressed, and annoyed wereyou?

Very Low Very High

Dr. Debayan Dhar
Department of Design

Now, what you see in this slide is an example of the task load index questionnaire.

What you see? It focuses on mental demand, physical demand, temporal demand, performance effort and frustrations and based on the evaluations of your users the final task load index has been calculated. This comes to the end of our lecture on the last module for this course. After this you will be exposed to one of the case studies given by one of my PhD scholars on health care.

And that would demonstrate the steps the procedures that have been covered to ensure that a requirement that has been defined by the designer it is addressed based on this nature of the context and specific instructions which is highly contextual because in this case the healthcare is a very complex in context. So, you would realize the kind of decisions making, the kind of complex situations that the designer faces in addressing this requirements.

Thank you.