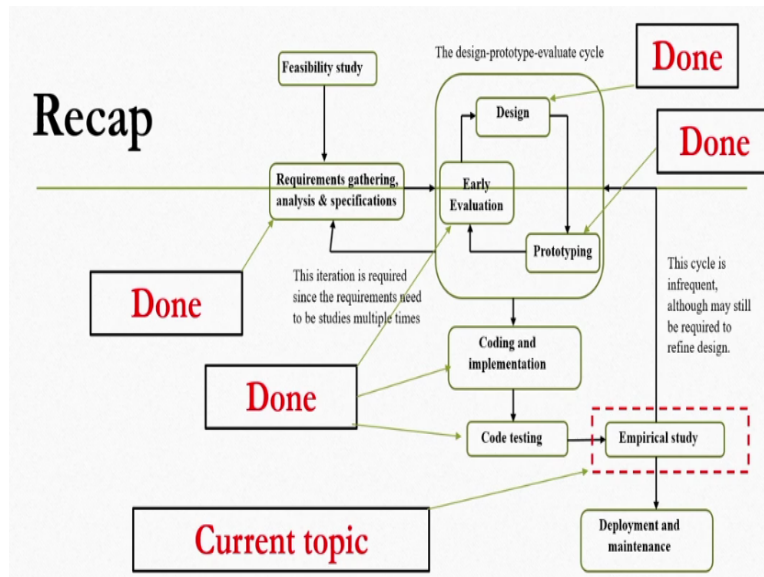


Design and Implementation of Human-Computer Interfaces
Dr. Samit Bhattacharya
Department of Computer Science and Engineering
Indian Institute of Technology – Guwahati

Lecture – 38
Empirical Usability Evaluation - 2

(Refer Slide Time: 00:54)



Hello and welcome to the NPTEL MOOCS course on design and implementation of human-computer interfaces. We are going to start lecture number 32 where we will continue our discussion on empirical study. As is customary before we start, we will quickly recap what we have learned so far and then we will come to the topic of this lecture. So, we are discussing the interactive system development lifecycle.

This is a systematic approach to build interactive systems. Human-computer interfaces are nothing but interactive systems. And the objective of having a systematic approach is to ensure that the end product which in our case is an interactive software meets two conditions, it is executable as well as usable. In the earlier lectures, we have talked about the various stages of the lifecycle in details including case studies on the outcome of those stages.

These stages include requirement gathering, analysis and specification stage. Outcome of this stage is software requirements specification document or SRS which we have seen earlier in details. Next is the design stage, in this design stage we primarily concentrate on design of interfaces and interactions and outcome of this is a designed document. Now, the design

document is created based on the experience of the designers as well as some design guidelines.

Now, once the design is created it needs to be prototyped for quick evaluation. So, prototyping stage also we have discussed, outcome of this is the prototype. Then comes quick evaluation of prototypes which is typically done with experts. Now, the objective of this design prototype evaluation cycle is to ensure that we come up with end usable interface design. After we arrive at a stable design of interface and interaction, we go for system design which is part of this design stage.

In system design, we try to design the code that is going to be written for implementing the system. Now, here our primary objective is to go for a modular hierarchical design and we can follow either of two design approaches. One is procedural approach; other one is object-oriented approach. For procedural approach we can make use of DFD as a language to express our design, for object-oriented approach we can make use of UML as a language for expressing our design.

So, the outcome of the system design phase is a design document of the system. Whereas the outcome of the interface design phases the design document for the interface. After the system is designed, we go for implementing the system, so we come to the coding and implementation stage. In the coding and implementation stage, we follow coding standards and guidelines to implement the system by writing programs.

So, the outcome is obviously the code itself along with some documentation of the code for better understanding. Once the code is written, we need to evaluate it for bugs. Now, evaluation can be done in different ways. A quick evaluation method is the review-based evaluation of the code where either of the two approaches or both can be deployed, namely inspection-based code review and walkthrough-based code review.

Outcome is as obvious a testing report. Now, the quick evaluation is one approach, also we can go for rigorous evaluation, more formal rigorous evaluation of the code following a functional approach which is known as black box testing and structural approach which is known as white box testing. In black box testing we do not bother about the internal structure

of the functions, rather we consider them to be black boxes and our only knowledge about those functions are what inputs they take and what outputs they produce.

Based on that we test the code that is black box testing and at the end we generate a testing report. In structural testing or white box testing, we are aware of the internal structure and we design test cases to execute the instructions or the statements that are present in the code. The outcome is again another testing report. So, testing is essentially a way to know about the code; whether the code is executable, what are the bugs and how we can overcome them.

Now, this testing does not say anything about the overall usability of the product. For that, we go to the next stage that is empirical study. Currently, we are discussing empirical study. In the previous lecture, we talked about few basic concepts of empirical study and started our discussion on different stages of empirical study. We are going to continue that discussion in this lecture as well.

(Refer Slide Time: 06:22)

Empirical Study

- Earlier, we learned how to evaluate usability of our designs
 - Done by experts, mostly (on prototypes)
 - In a limited scale

Now, as we have already mentioned earlier also in one of the stages, we talked about empirical evaluation, we talked about usability evaluation. First of all, why we need empirical study to understand usability of the end product? Now, earlier during design prototype evaluate cycle we talked about usability evaluation, but that was on a limited scale done by expert users on prototypes. So, that actually still do not reveal everything about usability issues as compared to getting it evaluated with actual end users.

(Refer Slide Time: 07:00)

Basic Idea

- We also need to **evaluate** our system with users, to get idea on usability

So, we need to evaluate our product for usability with end users and that too in a very systematic manner. We cannot go for evaluation in an ad hoc manner.

(Refer Slide Time: 07:12)

Basic Idea

- When we perform a **controlled experiment** to collect and analyze data on user behavior, the entire process is known as *empirical research*

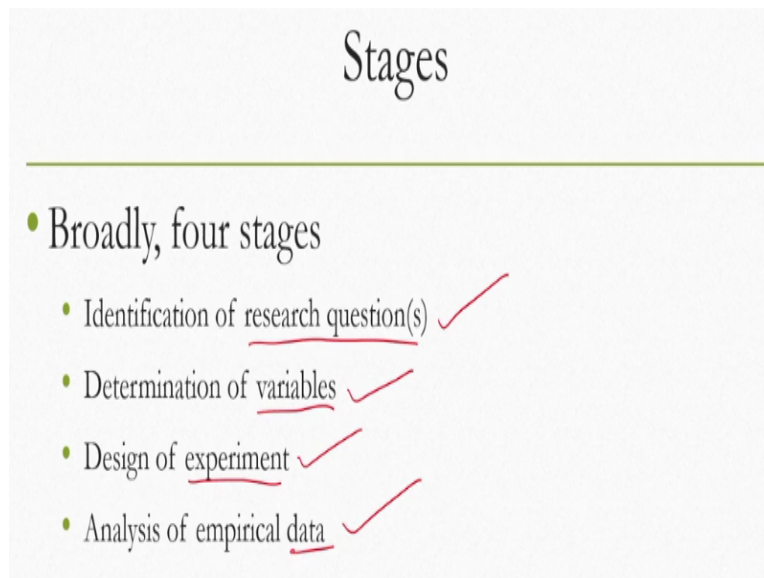
In order to perform systematic and rigorous usability evaluation of our product, we need to perform a controlled experiment, observe user behaviour during the experiment and from there need to conclude about usability of the product. So, this controlled study and observation of behaviour in that study is generally termed as empirical study, but this is not specific to usability evaluation only as we have noted earlier in the previous lecture.

This is a generic term where we observe and conclude based on observations. Here also, our primary objective is to observe user behaviour in a controlled environment that means we provide them tasks and we control the experimental conditions and under that controlled

environment, we observe how they behave while performing the tasks and based on that we conclude about usability by analysing the observed data that is the basic things that we have already discussed.

Also, we discussed in the previous lecture about the different stages of empirical research. So, empirical research or empirical study these terms we will use synonymously. In empirical study, we do several things and it is always useful to think of the study as consisting of distinct stages for better understanding of the process.

(Refer Slide Time: 08:51)



So, there are four stages that are there in any empirical study. What are those four stages? Identification of research questions that is the first stage, determination of variables second stage, design of experiment third stage and analysis of data that is fourth stage. Briefly, what these stages do? When we talk of empirical research, we start by asking questions. The whole objective of empirical research can be thought of as trying to find out answers to those questions.

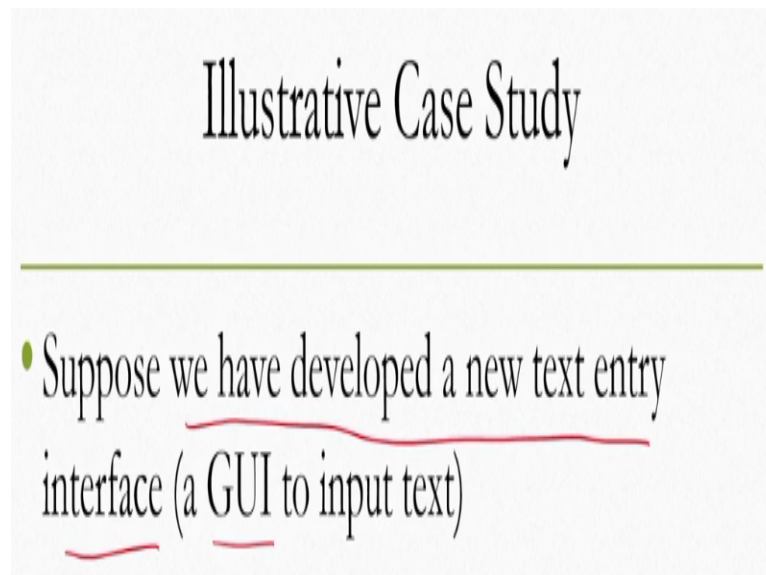
So, first step is we have to frame questions that is what we are calling as research questions. Good research questions are essential to perform a good empirical study. If the research questions are not properly framed, then the study results and the corresponding analysis of those results may not be very reliable. Once we are able to identify appropriate research questions, next thing is to identify what are the variables in our study.

Unless we are able to identify variables, we will not know what to observe and how to record those observations, so this is also very important. After that, we need to come up with a proper design of the experiment that we need to perform to collect empirical data or observed data. It requires planning, it requires careful considerations and balancing several trade offs so that design of experiment is our third step.

Once the experiment is done, we need to analyse the data, to come to a conclusion that is our fourth stage. One thing we need to remember which we already mentioned or already noted in the earlier lecture is that empirical study including all the stages are required to conclude reliably about the usability of a product. So, you may think of ignoring these stages, simply go to the users and ask for their feedback.

And based on that try to conclude something about the usability of a product, but that feedback may not be reliable. So, we require to collect reliable data to come to a reliable conclusion and for that we require a very systematic approach. And these four stages if done properly can provide us that systematic approach to come to a reliable conclusion about the usability of a product. So, in the previous discussion, we started our understanding of framing of research questions, let us continue that discussion in this lecture as well.

(Refer Slide Time: 11:40)



The slide features the title "Illustrative Case Study" in a black serif font at the top. Below the title is a horizontal green line. Underneath the line is a bullet point: "• Suppose we have developed a new text entry interface (a GUI to input text)". The words "text entry" and "GUI" are underlined in red, and "input text" is also underlined in red.

So, we took a case to understand the issue. So, the example case, example scenario was that we have developed a new text entry interface or a graphical user interface for text input and we want to understand its usability through empirical study. So, what kind of research questions we should frame to evaluate the usability?

(Refer Slide Time: 12:10)

Illustrative Case Study

- We wish to collect empirical data to measure its usability
- **First step – research question**

Note that our objective is to collect empirical data and to collect empirical data we must have a question based on which we will collect the data, so what should be that question? That is the first step, framing of the research question.

(Refer Slide Time: 12:22)

Research Question

- Consider the research question

RQ1: *Is the new technique good?*

In the previous lecture, we have seen that we can frame one research question which can come intuitively, too many of us; I am not saying all, but too many of us. The moment I say that you are supposed to frame a research question immediately this may be one of the questions that may come to our mind that is, is the new system or technique or the interface that we have developed is good? That seems to be a very intuitive question that comes to our mind. And based on this question, we have conducted an empirical study.

(Refer Slide Time: 13:09)

Research Question

- We present the **interface** to a user and ask him/her to judge its “quality”
- We repeat the process few times (say for five users) and complete our experiment

We have presented the interface to a user and asked for his or her opinion to judge its quality. Basically, we asked the users and we did this process for five times with five users. So, essentially in our empirical study, in our experiment what we did is we employed five users asked for their judgment on the quality of the interface because our question says that whether it is good; now good or bad is a quality. So, you ask the users for their judgment on the quality of the interface and recorded their responses.

(Refer Slide Time: 13:50)

Research Question (Recorded Observation)

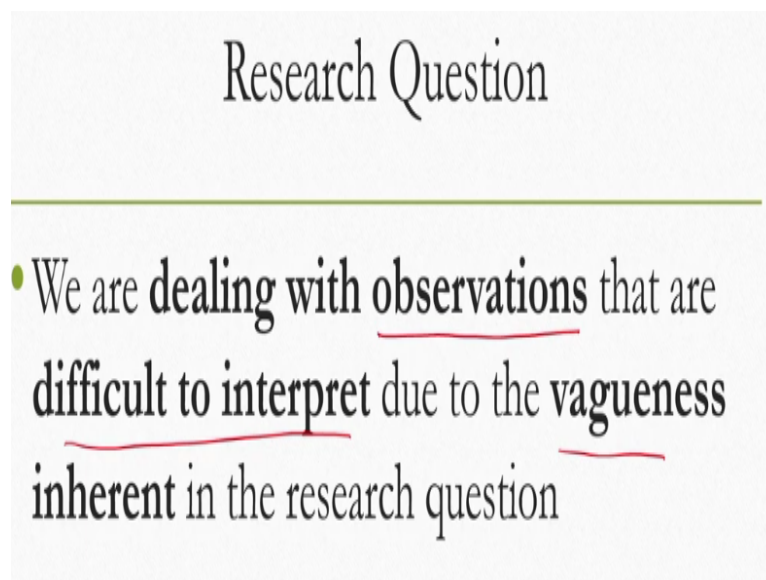
Experiment log	
Observations for our interface quality	
User #1:	good ✓
User #2:	poor ✓
User #3:	<u>not very good but not very bad either</u> ✓
User #4:	good ✓
User #5:	very bad ✓

Now, the responses that we recorded are like this which you have already noted earlier. User 1 said good, user 2 said poor, user 3 said not very good but not very bad either, user 4 said good and user 5 said very bad. So, here user 1 and 4 they have agreement that the interface is good. User 2 and 5 they disagree, according to them the interface is not good, it is either poor or very bad. Whereas user 3 did not give any definitive answer.

So according to that user it is neither good nor bad, somewhere in the middle. So, from these responses can we really conclude about the usability of the product? In fact, can we really conclude anything about the answer to the question whether our interface is good assuming that answer corresponds to the usability? That is very difficult because we do not have agreement, it is equally split between the users, the participants in the experiment, so very difficult to come to a conclusion.

In fact, it is impossible in this case, we cannot say with certainty whether we can call it good or bad. Now, why this problem happened? In spite of us framing a research question and asking for feedback, we did not get some observations which can lead us to definitive and reliable conclusion about the usability of the product, why this happened?

(Refer Slide Time: 15:34)



Research Question

- We are dealing with observations that are difficult to interpret due to the vagueness inherent in the research question

This happened because here we are dealing with observations that are difficult to interpret because we have framed a question which is vague. So, some vagueness is inherent in the question when we ask the users about their judgment on the quality where we explained the quality as either good or bad, we did not specify what is good, what is bad. So, users interpreted it in different ways and accordingly they gave the responses. So, the question itself is very vague and accordingly the answers; the feedback that we received were vague and difficult to interpret.

(Refer Slide Time: 16:16)

Research Question

- Let us now “compare” our interface with other similar (text input) interfaces

Can we do any better? Definitely we can and we have to do better to come to a reliable conclusion. So, let us see how we can improve on this situation, how we can come up with a better question. Let us know compare our interface with other similar interfaces, that means other text input interfaces. So, earlier we are not comparing anything, instead we were simply presenting the interface to the users and asking for their subjective opinion about the goodness of the interface.

We never said whether this interface is good with respect to so and so interface, so no comparison was made. Now, let us see if we compare the interfaces between our interface and some other already available similar text input interfaces, then whether we can do any better. So, first thing we have to do is we have to reformulate the research question.

(Refer Slide Time: 17:19)

Research Question

- We reformulate the research question

RQ2: Is our text input interface better than the text input interface provided by, say, MS Word?

So, let us frame another question, a revised question, let us call it RQ2 which is, is our text input interface better than the text input interface provided by some other interface, in this case MS Word, which is another text input interface Microsoft Word. So, here we are using the term better which is a competitive term. So, we are not using any definitive term like good or bad.

In comparison to some other interface whether our interface is better, so that is our research question. Now, if we conduct the same experiment that is ask 5 users for their feedback on this question, what kind of responses we are likely to get? Let us see.

(Refer Slide Time: 18:13)

The slide is titled "RQ2 - Observations" and contains a box labeled "Experiment log" with the following text:

- User #1: No. MS Word has many features.
- User #2: Yes. It is a clean interface.
- User #3: Difficult to say.
- User #4: Yes. Minimal useful feature set. Easy to remember.
- User #5: No. Can do many things with MS Word.

Now, we see somewhat different responses. User 1 said no, MS Word has many features. User 2 said yes, it is a clean interface; it means our interface. User 3 said, again middle ground, difficult to say; user 3 still not very sure about the interface. User 4 on the other hand said yes, minimal useful feature set, easy to remember. User 5 on the other hand said no, can do, it should be do.

Let us rectify it to remove ambiguity it should be do; can do many things with MS Word. So, are we in any better position? Apparently not because again here there is no agreement. User 1 and user 5 still agree that it is not better than MS Word. User 2 on the other hand and user 4 agree that it is better than MS word whereas user 3 is still noncommittal, still no definitive answer from user 3.

(Refer Slide Time: 19:33)

Research Question

- From these observations, we can conclude the following
 - When asked to compare, users give somewhat more concrete feedback
 - “Better” to some means more features, to others means less features
 - Some may still be confused (user #3)

Now, from these observations, what kind of conclusion we can make? When asked to compare, users give somewhat more concrete feedback. So, it is now in a better position to know exactly what is going on in the mind of the users who are giving the feedback. Better to some means more features, to others means less features. Note that here when we say is it better, some may think that better in the sense of having more features, so in that sense their answer will be no because MS Word provides more features.

Whereas others may think that it is a minimalist design, better means minimalist design, less features, so in that case they will answer yes because our interface is purportedly having less features than the MS Word. Some may still be confused, user 3, that is still possible. So, the takeaways are firstly somewhat more definitive, more concrete responses compared to the previous case.

But still there is ambiguity about the meaning of the word better and some users may still give confusing response. So, at the end we can say that we are in no better position than the earlier case, we cannot still draw any definitive conclusion which is reliable based on these feedbacks. But we get somewhat better responses compared to the previous one. Can we do even better? Yes, definitely we can do even better.

(Refer Slide Time: 21:16)

Research Question

RQ3: Does the new interface let me enter text
"faster" than MS Word?

Let us reframe the question, third revision of the question. So, earlier we have seen RQ1 and RQ2, now let us go to RQ3 a revised version of the question. Does the new interface let me enter text faster than MS word? Note how the questions are changing. Initially, we asked whether the interface is good irrespective of anything else. Next, we asked whether it is better than some other interface.

Now, in this third revised version of the research question what we are asking is a more concrete measure that is whether the interface let me enter or input text in a faster manner as compared to MS Word that is the other interface with which we are comparing.

(Refer Slide Time: 22:11)

Task & Experiment

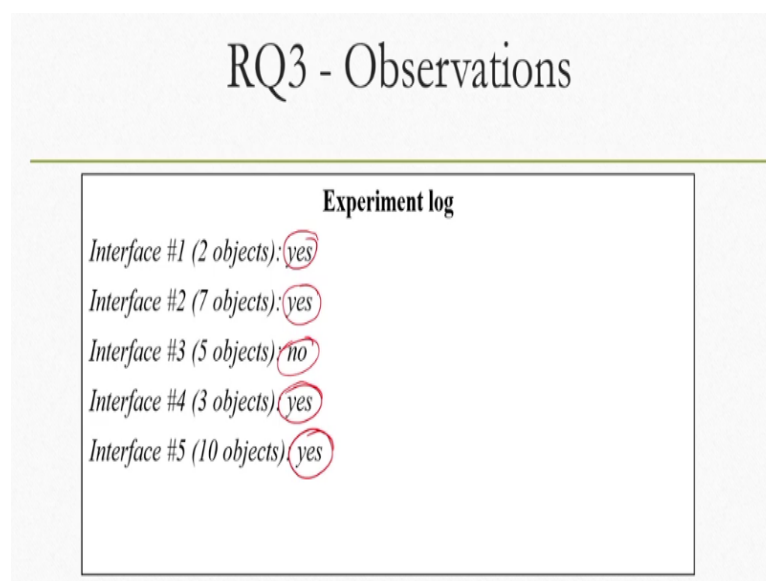
- Now, we don't just ask for feedback (which may be vague and subject to interpretation)
- Now, we can give them some task (input a text) and ask for their feedback

Now, in this case of course we do not need to take recourse to opinions or feedback of the users, we can actually measure the text input speed on the two interfaces and then decide

ourselves whether one is faster than the other. So, here we do not just ask for feedback. Now, why that is good because we have already seen that the feedback may be vague and subject to interpretation, so the feedback that comes from the users need not be in agreement.

Different users may interpret the research question differently and accordingly provide feedback. So, it is subjective feedback based on individual interpretations and the feedback can be vague, it is difficult to conclude from the feedback. Instead, what we can do? We can give them some tasks and ask for their feedback. Now, the tasks are since we are dealing with text input interfaces the task can be input a text and then we can ask for feedback.

(Refer Slide Time: 23:29)



Experiment log		
Interface #1	(2 objects)	yes
Interface #2	(7 objects)	yes
Interface #3	(5 objects)	no
Interface #4	(3 objects)	yes
Interface #5	(10 objects)	yes

Now, they have been given the task and they perform the task and then they are giving their feedback whether it is faster or not. Now, we are recording slightly more details. Along with the feedback, we are also recording number of objects present on the interface, however we will come to that later. So, for interface 1, user 1 said yes, it is faster. User 2 said yes, it is faster. User 3 said no, it is not faster. User 4 said yes, it is faster. And user 5 said yes, it is faster.

So, here what is the implication of this observation? If we ask whether our interface is faster than MS Word or any other comparable interface in a hypothetical case we are talking off and we ask the users to perform some tasks and then provide their feedback. We may expect that the feedback that we now get is likely to lead us to definitive conclusion.

(Refer Slide Time: 24:43)

Research Question

- Now easier to answer RQ3
- In fact, it's easier to analyze data avoiding subjective biases

It is now easier to answer RQ3 as we have seen and in fact it is easier to analyse the data avoiding subjective biases. Now there is no scope for individual interpretation, we have clearly stated that we are interested only in typing speed and typing speed can be measured in say words per minute or characters per second, we can measure the speed and then say whether speed on one interface is higher than the speed on other interface.

And then we can say it is faster than the other interface. So, no scope for ambiguity in the interpretation of the question. So, if we remove that ambiguity, then we get better observational data as we have seen with RQ3.

(Refer Slide Time: 25:36)

Research Question

- Suppose we gave them the same task (same text to input)
- We recorded text entry speed (number of characters entered per minute or CPM)
- We can compare the CPM values of the five users and come to an "objective" conclusion
 - Rather than relying on the users' subjective feedback on the idea of "faster"

Now, this can be done in another way as well. So, instead of asking for their feedback, we can do something on our own. Suppose, we gave them the same task. So, all of them perform the

same task that is the same text to input with the two interfaces, then we ourselves recorded the text entry speed in say CPM or characters per minute that is the unit of measurement for text entry speed.

Now, instead of asking for their feedback, what we can do is we can compare the CPM values of the 5 users and come to an objective conclusion. So, we no longer rely on their feedback, instead we can check ourselves the values because we have recorded those and then come to a conclusion about the speed of text entry on our interface in comparison to the other interface. So, rather than relying on the users' subjective feedback on the idea of faster, we can do it ourselves.

Although we said faster is less ambiguous compared to others, still faster to some can create confusion. So, instead of relying on their subjective interpretation which may lead to some sort of ambiguity in the outcome that is the feedback that they provide, we can simply remove the feedback component, we can record their speed by some means while they are performing the text entry tasks.

A simple way to do that of course can be suppose we have asked them to enter a 10 characters text, when they start entering, we start a stopwatch and then when they stop entering we stop the stopwatch. So, the time gap we can record and then we can easily compute characters per minute or characters per second from that time gap that we have recorded.

So, we do that for every user for every task and create a table and from there we can see their performance and we can say that given the tasks for user 1, one interface is faster than the other, for user 2 same interface is faster than the other and so on. So, this type of conclusion we can ourselves draw based on the observations that we have made. So, here there is no need for feedback which removes any possibility of ambiguity in interpretation of the research question. So, what it tells?

(Refer Slide Time: 28:32)

Research Question

- So, what is there in RQ3 that makes it **better** than RQ1
- TWO major differences

What is the difference between RQ1 and RQ2 and RQ3? What is there in RQ3 that makes it better than RQ1? I hope you realize that RQ3 is much better than RQ1 because it leads us or it can possibly lead us to a definitive conclusion to our research question in opposite to RQ1 where it is difficult to interpret the observations. So, RQ3 has two major differences as compared to the earlier questions that we have framed.

(Refer Slide Time: 29:10)

Research Question

- Lack of ambiguity
 - We are specifying that we wish to judge quality of interface in terms of “speed of text entry”

Lack of ambiguity. This is the first thing we should note. Here, we are specifying that we wish judge quality of interface in terms of speed of text entry. So, we are removing all sorts of ambiguity about what we mean by the quality of interface. So, we are avoiding the terms like good which is subject to interpretation, better which is again subject to interpretation and we are replacing those terms with a very specific term that is faster which is unambiguous.

And there is no possibility of different interpretations for this term. So, lack of ambiguity is the past characteristics of RQ3 which differentiates it from RQ1 and RQ2 that are the previous two research questions. There is a second crucial difference also.

(Refer Slide Time: 30:08)

Research Question

- **Measurable quantities**
 - We cannot measure interface quality – goodness is not measurable; it's always subjective
 - We have replaced unmeasurable concepts with measurable quantity – text entry speed

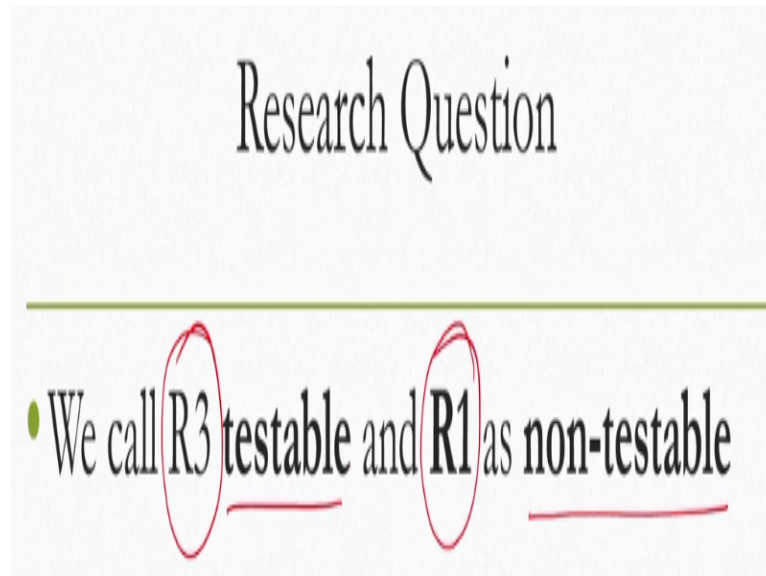
Now, we are dealing with measurable quantities, so this is very important. In first research question or second research question what we are observing, we are not actually observing anything that is measurable, instead we are collecting subjective feedback which we cannot measure. Whereas in case of research question 3, what we did? We defined a quantity faster. We introduced a quantity faster, which we can measure as I just mentioned.

How we can measure? We can simply record the time required to enter the text and then depending on the number of characters in the text and the total time required to enter so many characters, we can compute the speed and we can compare the speeds to know which one is faster. So, here all quantities that we are dealing with are measurable quantities, we can measure them and we can measure them objectively.

So, there is no scope or subjective interpretation of the quantities. We cannot measure interface quality like goodness or better, these are not measurable, it is always subjective. Whereas, in RQ3 we have replaced unmeasurable concepts such as goodness or betterment or better with measurable quantity which is text entry speed and using text entry speed, we can compute another measurable quantity that is faster.

We can simply compare, take the difference and see what is faster. So, both are measurable. So, we are replacing non-measurable quantities with measurable quantities in the third research question. So, these two are the crucial differences between the earlier questions and the third question that is lack of ambiguity and presence of measurable quantities.

(Refer Slide Time: 32:09)



Now, when we have measurable quantities in a research question, then we call it testable and when we do not have measurable quantities then we call it non-testable. So, R3 is testable, whereas R1 is non-testable. Now, we are introducing two terms, testable research questions and non-testable research question. Testable research questions are those which involve measurable quantities, non-testable research questions are those which do not involve any measurable quantities such as R1 and R2.

Now, there is a tradeoff between the type of research questions that we can have, whether it is testable or non-testable. Each has its own positive and negative sides and there is a tradeoff which needs to be balanced when we go for forming of research questions. So, let us try to understand the tradeoff.

(Refer Slide Time: 33:12)

Tradeoff

- Our aim - frame *testable* questions
- Problem
 - Testable questions designed to seek answer to *specific* queries
 - Such questions may lack *generalizability* to conclude about overall usability

What is our aim in empirical study? Ideally, our aim should be to frame testable research questions because this will lead us to reliable conclusions, definitive conclusions which is not dependent on subjective interpretations, so that should be our aim. Now, the problem with achieving this aim is that testable questions are designed to seek answers to specific queries, such questions may lack generalizability to conclude about overall usability.

Now, earlier we have seen that we asked for comparison in terms of whether one is faster than the other and that is a very specific thing. One thing we should notice that that specific question when asked leads to answer to those specific questions only. So, we can conclude that one is faster than the other, but then can that conclusion be generalized to say that one is more usable than the other or one is better than the other? That is not possible.

So, testable research questions by nature deals with specific queries which lead to specific answers and those answers lack generalizability.

(Refer Slide Time: 34:40)

Tradeoff

- Example – RQ3
 - Objective is to determine which interface is faster
 - Will that alone represent the interface quality?

So, we have already mentioned about this third research question RQ3. So, objective is to determine which interface is faster, will that alone represent the interface quality? So, here in terms of usability, we are more interested to know whether the interface is usable. But if we answered the question that one is faster than the other, can we conclude anything about the more generic concept of usability?

(Refer Slide Time: 35:04)

Tradeoff

- Text entry speed is NOT the only component that determines usability of an interface

Text entry speed which is the thing that we used to answer that particular question RQ3 is not the only component that determines usability of an interface. This should be obvious earlier we have talked about the concept of usability, it includes many things and only concluding about usability based on text entry speed and correspondingly whether something is faster than the other similar systems cannot lead us to a generic conclusion about the overall usability of the product.

(Refer Slide Time: 35:41)

Tradeoff

- There may be other aspects of the interface that determines usability
 - Number of features supported
 - Learnability
 - Error rate
 - ...

There may be other aspects of the interface that determines usability that includes number of features supported, learnability, error rate and so on. So, all these things we have already discussed in our earlier lectures when we talked about the idea of usability, so this is a problem. On the one hand, we aim to have testable research questions.

But then the problem is with testable research questions it may be very difficult to come to a generic conclusion about the overall usability of the products because testable questions lead to specific answers to specific queries. Now, those specific answers are not amenable to generalization about the overall quality or usability of the product.

(Refer Slide Time: 36:49)

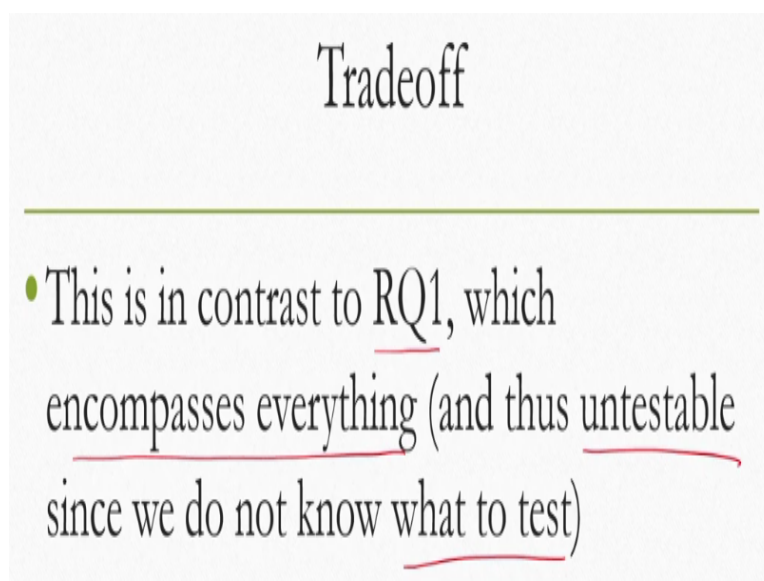
Tradeoff

- RQ3 answers only one aspect of overall usability
(efficiency)

For example, as we have seen RQ3 the third research question answers only one aspect of the overall usability that is the efficiency, how fast we can carry out the text input task, but it does not answer other qualities, memorability, error rate, learnability, satisfaction, etc. If we go by Nielsen's five measures or if we go by ISO standard definition of usability, it talks about efficiency, but it does not talk about effectiveness and satisfaction.

So, we cannot really conclude based on answer to R3 whether our interface is usable. We can say that our interface is faster than other interfaces, but we cannot say whether it is usable as compared to other interfaces.

(Refer Slide Time: 37:42)



This is in contrast to RQ1, the first question we framed which encompasses everything and thus untestable since we do not know what to test. Now, earlier we said RQ1 is not a good research question. But in another sense, it is a good question because it tries to come to a conclusion about the overall quality of the system that is whether it is good. But then since it tries to talk about everything, all aspects of the system, rather usability of the system.

It is untestable because we do not know then what to measure. There is no specific measure and it becomes very ambiguous to record observations. Based on what we will record observations we do not know, so it becomes untestable. Although the answer to this question can give us a generic answer to the broad question that is whether our product is usable.

(Refer Slide Time: 38:44)

Tradeoff

- However, if we can somehow get the answer to RQ1, we are supposed to get the true conclusion

However, if we can somehow get the answer to RQ1 we are supposed to get the true conclusion that is whether our product is usable. But to get the answer of RQ1 we need to test it which we cannot do because it is untestable, so that is the problem.

(Refer Slide Time: 39:04)

Tradeoff

- In scientific terminology, this is known as “validity” of the research question

In scientific terminology, this is known as validity of the research question. So, this situation that we are discussing relates to a concept called validity of the research question, whether it can answer generic questions or it can answer specific questions.

(Refer Slide Time: 39:25)

Tradeoff

- We can draw only specific conclusions from observations made for some research questions
- Such conclusions depend on the specific test conditions and generalization is not straightforward

We can draw only specific conclusions from observations made for some research questions. Such conclusions depend on the specific test conditions and generalization is not straightforward, in fact it can be impossible also as we have just seen.

(Refer Slide Time: 39:49)

Tradeoff

- The *extent* to which the observations made for a research question depends on the test condition is known as the “internal validity” of the question

So, the extent to which the observations made for a research question depends on the test condition is known as the internal validity of the question. So, if the observations made for a research question depends on the test conditions, the extent to which it depends on the test condition is called internal validity. That means the more dependent the observations are on the test conditions, we can say that it has more internal validity. And if it has more internal validity, then it is very difficult to generalize.

(Refer Slide Time: 40:38)

Tradeoff

- The extent to which we can generalize the conclusions drawn from the observations is called the “external validity” of the question

In contrast to that, the extent to which we can generalize the conclusions drawn from the observations is called the external validity of the question. So, the extent to which the observations are not dependent on the test conditions is called external validity. So, the more external validity that questions have, the more possible it is to generalize the conclusion drawn from those observations. So, what is the trade off?

(Refer Slide Time: 41:12)

Tradeoff

- A trade-off
 - We cannot frame questions that are based on generalized concepts (e.g. RQ1) - those are likely to be untestable
 - If we go for more specific questions (i.e., RQ3), we might get testable questions - however, we may not get the true answer

We cannot frame questions that are based on generalized concepts such as RQ1, those are likely to be untestable. If we go for more specific questions such as RQ3, we might get testable questions, however we may not get the true answer that is the tradeoff. We want true answer, for that we need to frame questions but those are not testable. Now, if we frame testable questions, we might get some reliable conclusions but that is not what we want at the end. So, there is a tradeoff between internal validity and external validity.

(Refer Slide Time: 41:50)

Tradeoff

- We can balance the trade-off by framing multiple testable questions

So, the research questions that we should frame should balance between internal and external validity. The more internal validity a question has the less generalizability, the more external validity a question has the less testability is and we need to balance the two, how we do that? We can balance the trade off by framing multiple testable questions. This is one way out instead of having a single question or limited number of questions.

We can set multiple testable questions, so multiple questions with high internal validity and the conclusions for these multiple questions can be together used to give us a conclusion about the overall usability or overall quality of the product. So, that is one way out and that is the way to balance the tradeoff.

(Refer Slide Time: 42:58)

Tradeoff

- Let us frame few more questions for our example

So, let us try to understand that with a few more questions that we can frame for our example.

(Refer Slide Time: 43:05)

Tradeoff

RQ4: Is the error rate within one hour of use less in our interface as compared to MS Word?

RQ5: Does the number of features supported in our interface sufficient to perform common tasks (provide feedback on a rating scale of 1 to 5 where 1 indicates "not at all" and 5 indicates "totally agree")?

RQ6: Can you remember the features easily (on a rating scale of 1 to 5 where 1 indicates "not at all" and 5 indicates "totally agree")?

Let us frame another question RQ4, is the error rate within 1 hour of use less in our interface as compared to MS Word? That is another question, which is of course testable. Another question, testable question is does the number of features supported in our interface sufficient to perform common tasks? In this case, we can ask the participants to provide feedback on a rating scale of one to five, where 1 indicates not at all and 5 indicates totally agree.

Yet another testable question, can you remember the features easily on a rating scale of 1 to 5 where 1 indicates not at all and 5 indicates totally agree. Why we are calling it testable? Because we are now collecting feedback on scales, it is not subjective opinions rather it is a quantitative sort of feedback collected on rating scales.

(Refer Slide Time: 44:16)

Tradeoff

- We captured different aspects of usability in the questions (on error rates, subjective satisfaction and memorability)
- We now perform empirical research for each separately
- Observations will lead to conclusion on overall usability - not possible with any one of the research questions

Now, with these questions are RQ3, 4, 5, 6 we can capture different aspects of usability of the system our interface, error rates, subjective satisfaction and memorability along with speed. Now for each of these questions, we perform empirical research separately. Observations made for each of these questions can lead to conclusion on the overall usability that is whether the interface is good.

So, if we find that our system is faster, more satisfying, easier to remember, less error rate then we can say that it is definitely good or it is better than other systems in terms of usability. So, we can come to that conclusion based on the conclusions we have drawn on these individual research questions. So, this is not possible with any one of the research questions.

So, we need to frame multiple questions, conduct experiments for each of these questions, draw conclusions for each of these questions and from those conclusions we can draw conclusion on the overall question. So, we have multiple testable questions and with high internal validity and for those questions we conduct experiments, perform empirical study and draw conclusions.

And based on those conclusions, we draw conclusion for a generic question that is whether our product is usable, whether it is good, whether it is better whichever is the generic question which has high external validity. That is the way out, that is the way to balance the tradeoff between internal and external validity of research questions between testable and non-testable research questions.

(Refer Slide Time: 46:17)

Tradeoff

- There is a positive correlation between the testable questions and the untestable question
- We are likely to arrive at a generalized answer for an untestable question from the specific answers to multiple testable research questions

Now, one thing is there. There is a positive correlation which seems to be obvious between the testable questions and the untestable questions. We are likely to arrive at a generalized answer for an untestable question from the specific answers to multiple testable research questions, this correlation probably exists.

(Refer Slide Time: 46:43)

Tradeoff

- A better approach than having only untestable question and user feedback

Because that correlation exists, we can then follow this approach where we can have multiple testable questions and based on that we can draw a conclusion to the untestable question. So, this is better than having only untestable question and user feedback which we have seen earlier.

(Refer Slide Time: 47:09)

Basic Idea

- “Testable research questions” are more popularly known as “research hypothesis” in the domain of behavioral research

So, we will end this with a small note on the idea of hypothesis. So, what is hypothesis and how it is related to empirical study because that terminology will be useful while you perform empirical study and analyse the data. Testable research questions are more popularly known as research hypotheses in the domain of behavioural research. So, they are similar in meaning, testable research questions are generally called hypotheses, but in a different form.

(Refer Slide Time: 47:39)

Basic Idea

- We start with two hypotheses: null hypothesis and alternative hypothesis
- Both originate from same testable research question

We start our empirical study with two hypotheses, null hypothesis and alternative hypothesis. Both originate from the same testable research question. So, we have one testable question which can give rise to two hypotheses; one is called null hypothesis, other one is called alternative hypothesis.

(Refer Slide Time: 47:56)

Example

- We can frame two *hypotheses* from RQ3
- H₀: Our design is not faster than MS Word.
- H₁: Our design is faster than MS Word.

For example, let us consider RQ3 that is our system is faster, is our system faster than MS Word? From there we can frame two hypotheses, one is called H₀, other one is H₁, H₀ is called null hypothesis, H₁ is the alternative hypothesis. So, in H₀ we frame it as our design is not faster than MS Word and H₁ we frame it as our design is faster than MS Word. So, what this tells us?

(Refer Slide Time: 48:28)

Example

- We are no longer posing any question - '?' at the end is gone
- Apart from that, there is one important difference - a single question gave rise to two hypotheses

One thing is we are no longer posing any questions, so question mark at the end of the research question is gone, so it is no longer a question. Secondly, there is one important difference that is a single question research question gave rise to two hypotheses. So, earlier we are dealing with one research question, now we are dealing with two hypotheses which originated from the same question.

(Refer Slide Time: 48:55)

Example

- In H_0 called the **null hypothesis**, we are essentially stating that the test condition is not going to affect the outcome (judgment)
 - Typically, opposite to what we set out to establish (effect of the test condition on the observations)
- H_1 called the **alternative hypothesis**, is just the opposite - we are stating that test condition does affect outcome

In the null hypothesis denoted by H_0 , we are essentially stating that the test condition is not going to affect the outcome or in this case our judgment. So, the null hypothesis typically states opposite to what we set out to establish that is effect of the test condition on the observations, this is what we set out to establish and null hypothesis typically states the opposite to that.

In contrast, the alternative hypothesis denoted by H_1 which state just the opposite. We are stating that the test condition does affect outcome. So, then what is our goal for any empirical study?

(Refer Slide Time: 49:48)

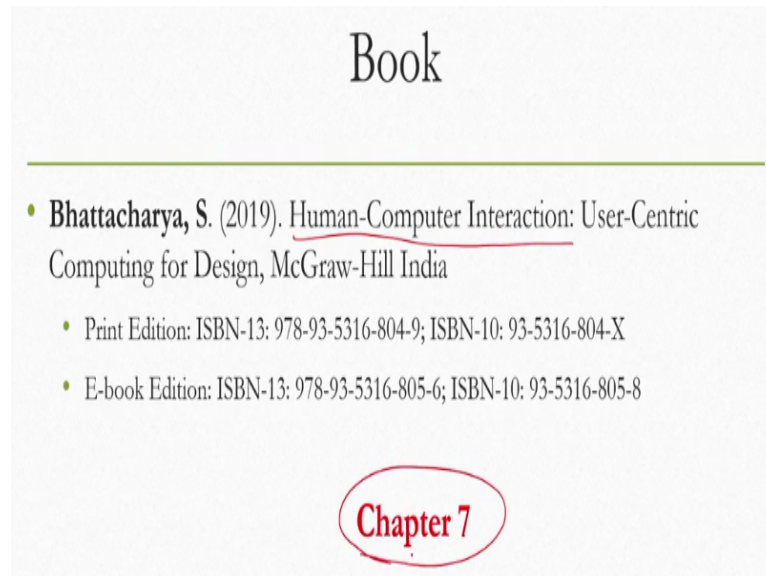
Note

- In an empirical research, we aim to find statistical evidence to refute or nullify null hypothesis and support alternative hypothesis

In an empirical study or alternatively empirical research, what we try to do? We aim to find statistical evidence, now this is a very crucial time to note, statistical evidence to refute or

nullify the null hypothesis and support alternative hypothesis. So, the objective of any empirical study is to find statistical evidence to refute or nullify null hypothesis and support alternative hypothesis. Now statistical evidence is not the same as simple conclusion. It requires proper analysis of data which we shall cover in a later lecture.

(Refer Slide Time: 50:33)



So, with that we have come to the end of this lecture. Here we talked about research questions, the issues, nature of the question, testable versus non-testable, the tradeoff which depends on the validity of the question. So, on the one hand we require questions with high internal validity to make them testable, but on the other hand our overall objective is to answer questions that are more generic having high external validity.

So, how to do that we discussed. The idea that we discussed is that we frame multiple testable research questions to come to a conclusion about more generic non-testable question. And at the end, we briefly talked about the idea of hypothesis, so that is similar to the concept of a research question, but instead of research question which is a question as the name suggests, hypotheses are not questions.

We have two hypotheses from one research question, one is null, one is alternative hypothesis. Our objective in any research, empirical research or empirical study is to refute null hypothesis and support alternative hypothesis with statistical evidence which is a special data analysis technique that we shall learn in a later lecture. I hope you enjoyed the lecture and understood the concepts.

More about this topic can be found in this book Human Computer Interaction chapter 7. Looking forward to meet you all in the next lecture where we will continue our discussion on the other stages of the empirical study. Thank you and goodbye.