Human Computer Interaction (In English)

Prof. Rajiv Ratn Shah

Department of Computer Science and Engineering
Institute IIT Madras

Evaluation

Hello, everyone. Welcome back to the class on human-computer interaction. So this week, we will be talking about evaluation. Why is evaluation important? How do we know that the system we have developed is fulfilling the needs of the users and serving the purpose for which it was designed? So before that, let's do a quick recap of what we have covered in the previous lecture. So we have studied what prototyping is? What is prototyping in interaction design? We have discussed a lot about the smart UI, how it is useful, and how it changes the world. Specifically, it is changing companies like Tesla, SpaceX, and so on.

becoming billion-dollar companies. So we have also discussed the tutorial and a panel discussion on prototyping and the assignment around that. So this week we'll be talking about what an evaluation is. What is experimental design? What are the types of evaluation? What are the evaluation methods and metrics? Through our favorite use case, the Blade Learning App, we'll see how to do an evaluation.

Finally, we'll do a tutorial on evaluation with hands-on experience using Adobe Illustrator. So, what is an evaluation? Evaluation in SCI refers to assessing the usability, effectiveness, and user experience in an interface or a system. So I believe you recall what usability is. How users are able to achieve their goals in an efficient, easy, and memorable way. So we discussed a lot about usability in our earlier lectures.

We'll do a recap as well: what it is and what we are willing to achieve. So it also helps to measure how well a product meets the needs of users. and how users interact with it, ensuring that the product is intuitive, efficient, and satisfying. Because if the user is not satisfied, the user is going to have a rather low user experience. And if a user is showing less user experience, in a way, they are not going to use the product that you have built for them.

So, as Don Norman rightly said, evaluation is not a luxury; it is a necessity. Why is it a necessity? Because when any company comes up with a solution or a product, they can create a lot of hype. Would you buy a product or use a service just based on the hype? I

believe not, because there should be some accountability. There should be some justification, and there should be some reasoning behind any claim that the company is making. And that can only be done through a thorough evaluation.

And can you guess what will happen if a thorough evaluation does not occur before launching the product? What if you start building a product that probably does not align with the user needs? What if you build something that the user is probably not looking for? So evaluation is the key where you try to match what the system is that you have built. What the user is looking for, what the system is designed for, and whether the system is actually completing the task. It has to. So there are bigger challenges and major issues if you don't do a thorough evaluation. So let's see an example.

Google's new launch board made the company lose $100 billion in just one day. Can you imagine how big this number is? How big is this amount? And it happens because in just being the race to launch LLMs just like OpenAI, They have just launched BARD in such a hurry. That they could not evaluate all possible test cases. scenarios and that is the reason during the launch basically it has committed a mistake which it's not supposed to do and that has basically cause the Google share plunges and just losing $100 billion in just one day. Another very popular example, probably you may recall, CyberCab, where Elon Musk launches it.

And he tried to show that, I mean, even if you hit by a hammer on the windows, the glass won't broken. And what happens? The rest is history. So in his demonstration, when he hit the car window by hammer, it has broken down. In a way, the thorough testing has not been done and that has cost just $15 billion and so on. Same with the case with many other companies, be it like Meta, be it Microsoft, Snatch, so on.

With the different conditions, with the different use cases, with different evaluation, even they did not probably understand the user need properly and probably even there they do understand the need and the probably requirement they probably they haven't built something which is as per the expectation of the users and that's where basically they lose money they lose reputation and so on so that is the reason we it is very important to do a thorough evaluation for a product that you are going to launch. And these evaluations should not just be done on some simulated data or on some simulated users or some dummy users. Ideally, these evaluations should be done in the natural setting. That's what I say. You need to get your shoes dirty.

You need to go into the you need to go into the shoes of real users and see whether they are happy with using the system, are they able to complete the task they're supposed to do with fun, enjoyment and without any hassle and so on. So that is highlights the need of a

thorough evaluation especially in real environment where the product is going to be used with probably the real set of users that you are going to have. So I believe you recall the usability that we discussed in a couple of earlier lectures. So usability in HCI refers to the ease, efficiency, and satisfaction with which users can achieve their goals when interacting with a system, product, or interface. It is a key aspect of user-centered design and focuses on making system intuitive, accessible, and effective for the intended audience that you have.

So usability is critical for ensuring system meet user needs effectively, enhancing adoption, retention, and overall user satisfaction. So if usability criteria is not met then it will lead to rather than adoption people start losing it people probably start moving to some other similar system a better system and overall user will not be satisfied so we have to ensure that when a user or any stockholders interacting with the system or the product that you have built, it should be easy to use, it should be efficient, and users should have satisfaction after using it and feel that the sense of achievement, what they were planning to achieve, they were able to do very efficiently, effectively, and in a simple and easy way. So there are several key dimensions of the usability. And if you recall, we discussed effectiveness, efficiency, learnability, memorability, error tolerance, and satisfaction. So in order to make sure that your app is meeting the usability criteria, we have to work on all these.

Effectiveness primarily related to the accuracy and completeness with which the user achieve their goals. So system should be effective. For instance, whatever the functionality it's supposed to achieve, it should complete the sale. For example, if you're using a digital payment app, say Paytm, PhonePay or whatever, if you have to transfer say 100 rupees to some person x then ideally it should suppose to transfer 100 rupees not 99 or 101 or 1000 and it should not get failed so effectiveness is primarily about the accuracy and the completeness with which user achieve their goals efficiency primarily about the speed and the ease of achieving the goals with minimal effort and time so there could be different metric you can talk about so for example let's consider the metric of number of clicks to achieve the things so for example with how many minimum number of clicks you can transfer 100 rupees to your friend if the app for example paytm is able to do that in just two clicks and another counterpart app probably is doing in the same transfer in say three or five clicks, it means the other app is not efficient as it should be. Another criteria of efficiency could be the time.

For example, for transferring 100 rupees, the app X is taking say 100 milliseconds, but the app B is taking just 50 milliseconds. then definitely app B is more efficient than app A. Similarly, you can think of more metrics and the constraint through which you can define the efficiency. Another thing to discuss is learnability. How easily new user can

accomplish tasks on their first attempt.

If you are able to do that, quite easily and especially if your system is designed in such a way that it can somehow use the common sense of the user. For example, you may have a user who might be using Paytm. What if you ask the user to use, say, Phone Pay or Google Pay or so on? Ideally, the new app should be able to use the common sense knowledge of the user to use your app and probably perform the same task with very minimal effort. So how easy, how easily users can accomplish the task on their first attempt. So that is something about learnability.

Memorability, how easily returning user can re-establish proficiency. For example, you are using Paytm, you probably use for five years and then you move to America where probably the Paytm is not working and you there you stayed there for another five years how memorable the Paytm app is so for example you come to India for a holiday and if you are If you have to use PTM, how easy for you to memorize it and probably use PTM again without any hassle. So that is about memorability. Error tolerance, how will the system prevent errors and help user recovering from them? So it's very common that probably user can commit a mistake. even the some error can happen due to system Wi-Fi connection internet connection and so on or probably usually the resolution in the mobile is quite small so by mistake probably you hit some other button instead of the button that you supposed to be so how do you recover from recover from these mistakes or errors so that is about error tolerance.

And finally, the satisfaction the users come for and overall positive experience while interacting with the system. So ideally it should be as high as possible to retain the user, to keep the user happy and provide a sense of the user has completed the task with very minimal effort and it's effective, it's efficient, it's learnable, it's memorable, error-tolerant and so on. So real-world example you can think of a ride-hailing app in India like Ola and Uber should allow their user to book ride quickly. So that is about efficiency. Whenever you click book button, ideally you should be able to allow to book a cab.

So that is more about effectiveness. Minimal errors such as incorrect pickup location. So often you see that when you select current location, so probably it select the next block next to your house and that's mistake. So how to correct that? How to recover that? That's more about error tolerance. Be intuitive enough for the first user to navigate easily, that is learnability.

So even though I might be using Uber before, what if I have to now start using Ola? Is it easy for me to just learn it and go ahead? And even though I've not used any app before, how easy for me to basically just learn it very quickly and just get started? Similarly,

ensure users feel satisfied with the experience, including payment and feedback process, is about satisfaction if the overall satisfaction about the booking that you have it's totally dependent on were the pickup was on time were the driver has taken probably right amount of money did not charge for extra money probably did not force you to probably get out of the car even just before or probably some other drop location and so on so there are minimum cases there are several cases you can think of so there are several cases around this you can think of so an activity for you So can you do same thing for effectiveness, for efficiency, for error tolerance, learnability and satisfaction about a live streaming app like Jio, Hotstar or probably Disney and so on. So try to think about it, different use cases, why and how these usability things are made. So another activity. Find the list of products which are failed due to lack of thorough evaluation and user study. earlier slides we discussed the importance of evaluation why it is important and if a thorough evaluation has not been done then it will lead to the failure of the product or probably it is going to put a huge amount of monetary loss for you.

Google Bard's example we have seen where it was launched without thorough evaluation and that's where basically it failed and it costed $100 billion loss in just one day and so on. So think something around your real life and see can you list down those such products which are failed due to thorough literature survey, which are failed due to thorough evaluation, which are failed due to thorough user study So think what you would have done to avoid the failure if you would have been the CEO of the company. So think about it. Similarly, find the list of products around you in real life and tell what motivate you to use it or probably leave it. You continue using the app probably once the app fulfill the usability criteria like you are able to achieve your goals with easy efficient and effective way it easy to learn and easy to memorize and it can further help you in recovering from the any errors that has been made along the process be it by you or through the system, through the app.

So if you're satisfied, then probably you can continue using it. If you're not satisfied, if you're frustrated, you're going to leave it. And doesn't matter how good the functionality it has, if it is not easy to use, if it is not easy to remember, if it is not easy to learn, you are most likely not going to continue with it. So think about such products, services around your real life. Let's go for evaluation in human-computer interaction.

Iterative design and evaluation is a continuous process that examines these four Ws, why, what, where, and when. Why is about to check user's requirement and confirm that users can utilize the product and that they like it what is about a conceptual model early and subsequent prototypes of a new system more complete prototypes and a prototype to compare with the competitors product and where ideally it should be in the natural in the wild and or in the laboratory settings When? Throughout the design because it is more

iterative in nature. Throughout design, finished product can be evaluated to collect information to inform new products. So what are the benefits of evaluation in SCI? It improves the usability. Regular evaluation helps identify usability issues early in the design process.

and what will happen if you identify the issue in early design process. You don't have to redo the things. You don't have to waste a lot of your time, resources in building what you are doing. to enhance user experience. So it ensure that the design meets user expectations and deliver a positive experience.

And positive experience will lead to satisfaction, a better user experience. Confirm design decisions. It helps validate whether the design choice align with user need and context or not. If it doesn't align with user needs or the context, it's not going to work. The user is not going to adopt it or use it or pay for it.

Minimal risk. By testing with real users, you can minimize the risk of building a product that doesn't work well in real-life scenarios. Since you already involved your users in early stages, you are taking the regular feedback with them. They are trying your system, and in those cases, even they tend to forgive you as well if there are some little mistakes. Better user engagement. So regular evolution help ensure that the system align with the user's goal and preferences, resulting in better engagement.

It also helps in reducing the cost by identifying and fixing usability problems early. Evaluation reduces the risk of costly redesign after launch. It is also useful in providing insight for continuous improvement. Ongoing evaluation helps refine and improve the user experience over time. So for example, Smartphone app evaluation testing a new social media app with users to ensure that features are easy to use and the interface is intuitive, simple, and so on.

So evolution is not a one-time event. As I said earlier as well, it is iterative in nature. It has to be performed in different stages, many times with different users, different stakeholders, in different environment. It is an ongoing process that ensure that the design remains relevant and useful, Don Norman said. So, we got a fair understanding of what is an evaluation so let's see what is experiment design so in experiment design we study what is the control settings that directly involve users so for example usability or research labs natural settings so that involves user for instance online communities and products that are used in public places Often, there is little or no control in the natural settings what users do, especially in the wild settings.

They can do anything what they like. They can play with the system any way they can.

Any setting that doesn't directly involve users, for example, consultant and researchers critique the prototypes and may predict the model how successful they will be when used by the users. so in the natural setting that is more about you can say the field study so where field studies are done in the natural setting and it is also known as the in the wild is a term for prototyping being used freely in the natural setting where in ideal world it will be used by the real users seek to understand what user do naturally and how technology impact them Field studies are used in product design to identify opportunity for new technology, determine design requirements, decide how best to introduce new technology and evaluate technology in use. So in the experiment design, we also talk about who are all the users we have. what are the different participant you have so it could be between subjects so single groups of participant is allocated randomly to the experiment condition so the benefit is no other effect but the drawback is individual differences cause the different results same participants within subjects group so all participants appear in both conditions The benefit is few and no individual differences, but the drawback is counterbalancing need due to ordering effect.

It could be match participants, pairwise, where participants are matched in pairs, for example, based on expertise, gender, or so on, you can match the participant. So benefit is individual differences reduce because now they are matched based on different similarity. And the problem is they cannot be sure of perfect matching on all differences. You don't know. In experiment design, another factor which matters, how many evaluator we need to have it? So Nelson suggests that on average, five evaluators identify 70 to 80% of the usability problem that you have.

There are other work as well, which probably also point out that it also depends on the context and the nature of the task, but you can assume that in most of the cases, five to six evaluators are enough to identify 70 to 80% of the usability issues. So what are the evaluation type you can follow? so primarily there are two types of evaluation but there are additional one as well that i will explain but most popular are two the first one is formative evaluation so conducted during this design process to gather feedback and guide the development of the interface So it helps refining prototypes and design ideas. So example is user feedback from a prototype of a navigation app is gathered before the final version is developed. And the benefit is it help inform design decisions and provide early stage feedback before the launch.

So there are different methods has been used. So usability testing, expert reviews, that is more about heuristic evaluation, survey and questionnaire, that is more pre-release feedback and so on. The second most popular type of evaluation is the summative evaluation. So it occurs after the product is developed to assess its overall effectiveness and usability. So typically involves usability testing with a larger group of users. For

example, post-law survey for a website to measure user satisfaction and gather feedback on the final product.

The benefit is provide final confirmation on the product's usability, because it is primarily after the launch, and helps assess user satisfaction after using the full product. And there are different methods for this, post-launch surveys, A-B testing, final usability testing, and so on. There are other type of violation also happens, but from different perspective. The third one is called diagnostic evaluation. So it focuses on identifying specific problems or issue within a system that is more about finding the health of the system.

So it helps to understand whether the usability issues are occurring in the design. For example, using cognitive walkthrough on an online booking system to identify issues in navigation and task completion. So benefit is it helps pinpoint specific usability issues and allows designer to make targeted improvements. So methods used for this is task analysis, cognitive walkthrough and heuristic evaluation.

So next type of evolution is called longitudinal evolution. It involves studying user interaction over a long period to understand how the product performs and how users adapt it over time. So for example, conducting year-long study on the uses of fitness app, observing how users engagement and goal changes over the time. So the benefit is it tracks long-term user behavior, which is very important to note down, and satisfaction and identifies evolving needs and potential issues. So method used for this is user diary, long-term usability testing, and survey over time. So what are the evolution methods are? Some names we have discussed in earlier slides briefly.

So let's go about it. So first is usability testing. So it involves observing real user as they interact with the product to identify usability problems, task completion difficulties, and area for improvement. For example, testing a website redesign with users to see if they can easily navigate to key section that you are expecting the user to go and complete tasks like making a purchase. Are they able to purchase the product that you are willing to sell? So benefit is provide real user feedback and helps identify design flaws. Steps involved Define tasks for users Observe how they interact with the system Collect qualitative and quantitative data about the same evaluation method is heuristic evaluation so a usability inspection method where a small group of evaluators it could be expert it could be some identified stakeholders who assess the interface based on the established usability principle heuristics Example, an expert reviewer uses Nielsen heuristics that will come in a while to evaluate the usability of a shopping cart interface on an e-commerce website. The benefit is it's quick and cost effective and identify many usability issues early.

It is developed by Jacob Nielsen in the early 1990s. So steps involved, expert review the interface, They identify potential usability issue based on the heuristics. The results are compiled and prioritized. So this is the revised Nielsen's heuristics in 2014. So that focuses on recognize rather than recall. And if you see this is inspired by human way of Remembering the thing for human as well, recognition is easier as compared to recalling.

Flexibility and efficiency of use. Aesthetics and minimal design. Help user recognize, diagnose, recover from errors. Help and documentation. Visibility of system status.

Match between systems and real world. How close they are. user control and freedom, consistency and standard, error prevention, and so on. So during his heuristic evaluation, a briefing session to tell expert what to do, an evaluation period of one to two hours or even longer, each expert works separately, take one pass to get feel for the product, and take a second pass to focus on some specific features, and briefing session in which expert works together to prioritize the problem. So what are the major problems here? Few ethical and practical issues to consider because users not involved. It can be difficult and expensive to find experts.

Best experts have knowledge of application domain and users. And the biggest problem is important problems may get missed. Many trivial problems are often identified such as false alarm. Expert may have bias how to deal with that. So the next evaluation method is cognitive walkthrough. So here evaluators walk through the user interface step by step to identify potential issues in user's cognitive flow.

So for example, evaluating an online bank transfer process, analyzing whether users can easily understand and execute the steps for transferring money. the person is able to do easily with minimal number of clicks and effort and accurately we are good but if the user is facing any challenges in any steps then that is something we have to figure out that is something we have to work on so it helps uncover issues that might be missed in the other evaluation and focuses on the user's mental model during interaction. So steps involved, select specific task to evaluate, walk through the interface as a user and noting all the challenges, identify potential cognitive obstacle for the users. So designer presents an aspect of the design and user scenario. Expert is told the assumption about the user population, context of use and task details.

One or more experts walk through the design prototype with the scenario. An expert guided by the three questions. Will the correct action be sufficiently evident to the user? Will the user notice that the correct option is available? And the third one is, will the user associate and interpret the response from the action correctly? So as the expert walk

through the scenario, they note down the problem. And once you note down the problem, you can work on it, you can fix them. So the next method is survey and questionnaires. So you can collect feedback from a user through structured questions to assess their satisfaction and experience.

Sending a survey to a user's app to gather feedback on the app's usability and features. So benefit is it's cost effective way to gather data from large number of users. We also discuss how to collect data from a crowdsource scenarios so can be conducted remotely making it accessible to diverse set of users helps identify trends pain points and user expectations provides structured data that can be easily analyzed other ways you can think of is eye tracking that is more advanced so it involves using technology to track where and how a user look at different parts of an interface, providing insight into their attention and behavior. So for example, if you're placing a sale advertisement on e-commerce website, then ideally you're expecting the user to look at the sale and probably accordingly probably select the sale item, say a phone or washing machine or fridge, whatever, and buy it. If the user doesn't look at there, if the user doesn't click there, it means there is something is missing.

Similarly, using eye tracking to evaluate how users focus on elements in a news website layout and where they get confused or distracted. So often people say that, I mean, how to keep track of eye tracking? Not all the users are using eye tracker. How you will collect that data? So there is another research paper around that direction, which says that even you can simulate that where the person is looking at based on the different user behavior. So for example, where my mouse cursor is, where, so these mouse cursor basically a kind of one indicator of where I'm looking at. So this is, this may not be the perfect data, but it's a good amount of data and clean data that is give a very good overview of where the person is looking at so the benefit is provide precise data how user interact with the interface visually identifies area of attention distraction or confusion If at some point, for example, here I'm confused what to do, then probably I spend more time or probably look at there.

Otherwise, if it's clear to me, I'll just simply click and go ahead for the next page or probably add to the cart and probably buy the item. So another method is think aloud protocol. So where it involves asking users. So you basically ask user to verbalize their thoughts while interacting with the system.

It is one of the very effective way of evaluating the system. So it providing insight into their decision-making process. So for example, a user verbally explain their choice as they navigate through a video streaming service, helping identifying cognitive bottlenecks. So the benefit, it offers detailed insight into user behaviors and mental

process and help understand why user makes certain decisions. So the steps involved user complete task while speaking their thought aloud. So you let the user play with the system and in the meanwhile, you ask user just keep telling what is happening, what you're doing, what you're going to do next.

So researchers observe and analyze the reasoning behind the different user action which they have taken. So another method of evaluation is remote usability testing. So there might be the cases where you do not have access to the real users. So involving user testing the product in their natural environment while the researchers observe their interaction remotely. So if you want the user to use your app or system in the natural setting so that you can collect the data more appropriate data that should be considered for evaluation of the system or user behavior.

You have to let the user continue using the app in the real setting. and probably you can't be at all such places. In that case, remote usability testing can help. So for example, conducting remote usability testing of a home automation app where participants control smart devices in their own homes. So benefits, it allows the testing with the users in different locations and more natural, less artificial than lab-based testing. Another method of evaluation and one of the most important one as well, evaluating for accessibility using guidelines.

As we discussed in a couple of earlier lectures, there are several accessibility guidelines such as WCAG. And one of the most well-known analytics is that. So the next is predictive modeling. So it provides a way of evaluating products or design without directly involving users, hence less expensive than user testing.

So you can build a model which can do the task for you. Usefulness limited to system with predictable tasks, for example, voicemail systems, smartphones, or dedicated mobile phones. It is based on expert error-free behavior. So Fitts' law predicts that the time to point at an object using a device is a function of the distance from the target of the object and the size of the object. So Fitts' law say that, so if you want to point to an object, probably your current location is this. So if you want to go there, and also depends on the size of the object, is a function of that so if you can keep that in mind then probably you can take a call where the next button should be how big it should be so that you can take minimal amount of time to reach to the next logical button on steps and so on one simple example would be while filling the form you see where the next button is just and usually when you fill it you go to the next page and again, next, next, next, and so on.

So you can design that in better way by following the Fitt's law so that minimal number of navigation should be happen, minimal number of movement should be happen in order

to save precious time. The further away, the smaller the object, longer the time to locate it and point that's what i try to explain here it is particularly useful for determining where on a screen to position an object so fitz law is useful for evaluating system for which the time to locate an object is important for example smartphone handheld etc so that is about evaluation methods so let's talk about evaluation metrics So there are several evaluation metrics in HCI and evaluation metrics in HCI are essential for assessing the usability, effectiveness and overall user experience of the system. So these metrics basically help in understanding how will a product meet the user needs and expectation. So they are used in formative and summative evaluation that we discussed earlier to guide the design process and validate design decisions.

So Peter Drucker rightly said, what gets measured gets managed. Because if you can't measure, then you may expect some surprises when the user is going to use it. You don't know what is going to happen next. So as part of thorough evaluation, when you do the measurement, when you measure the system performance and everything, you know what would be the next expected outcome, what would be the next logical steps and so on, how the user is probably going to behave, how they are going to achieve answer in different situation, different scenarios. So the popular evaluation metrics are efficiency, effectiveness, and satisfaction that we discussed earlier as well to describe in more details.

So efficiency is about how quickly and accurately users can complete the task. So for example, measuring the time it takes a user to complete an online shopping checkout process. Effectiveness is how will the system support user in achieving their goals. So for example, evaluating how successfully user can find and select a flight on an airline website. Satisfaction is about user's perceived satisfaction with the product usability, design, and overall experience. So you can see example of using a system usability scale survey to evaluate satisfaction of a newly designed app.

So these are evaluation metrics you can think about. Usability metrics, user experience metrics, performance metrics, cognitive metrics, behavioral metrics, error metrics, biometric metrics, and social metrics. So let's talk about each of them one by one. So usability metric is about how To measure effectively and efficiently, a user can interact with the system. And the key metrics here are task, success rate, time on task, error rate, efficiency, and liability. So example is measuring the time it takes for a user to navigate through a banking app to complete a money transfer.

So that is more about usability. So about user experience, it evaluates the subjective experience of a user interacting with the product. So the key metrics here are satisfaction score, net promoter score, SUS, system usability scale, engagement, and emotional

responses. So example is using an SUS questionnaire to assess the usability of an e-learning platform. Where the satisfaction score is primarily about the overall rating given by the users for their experience.

NPS is primarily about the likelihood that users would recommend the product to others. Because you are satisfied, you think this is going to be useful to your friends, colleagues, and so on. SES is a standard questionnaire that provides a usability score, engagement frequency, and the duration of user interaction with the system. An emotional response is about users' feelings and emotions after using the product. If, after using the product, The feelings and emotions are positive; they are going to use it and are likely to recommend it to their friends and so on.

If it is bad, they're not going to use it. They're not going to come back. And even they can stop the other friends and colleagues from using it. Similarly, the performance metric focuses on how the system will perform in terms of speed, accuracy, and reliability. And the key metrics here are response time, system load time, accuracy, and downtime. So measuring the response time for a voice command in a smart home assistant.

Similarly, cognitive metrics are more about the mental effort required by the users to complete a task. Ideally, you want users to apply minimal mental effort. So the key metrics here are mental effort. amount of cognitive resources needed to complete a task. Recall accuracy, the ability to remember and use features after a period. Cognitive load is primarily about the mental effort required during interaction, often measured through eye tracking and subjective rating scales.

Time to recover is the time it takes for a user to regain focus after an error. Example is evaluating the cognitive load of medical professionals using complex patient management software. Similarly, regarding the behavioral metrics, it tracks observable actions of users to identify the interaction patterns. How does the user behave? How does the user use it? And so on. So the key metrics are click rate, navigation path, drop-off rate, and heat maps.

So the click rate is about the number of times a user clicks on specific elements. So, for example, for a given website, if the person keeps clicking help, help, help, help, it means there is something wrong. Again, this is kind of indicative. Navigation path: the path users take through the website or app. The drop-off rate primarily refers to the percentage of users who abandon the task at a certain point because they give up, think it is useless, or are unable to complete the task, and so on. Heat maps are the visual representation of user clicks, crawl, and attention, so you might have seen on the web that there are some areas which are red and bigger, some which are probably lighter, yellow, green, and so

on, so you can probably define the red.

The redder the area is, like a heat map, the more people have probably spent time there; for example, analyze user frequency while filling out an online form, noting where users make the most mistakes. Similarly, biometric data involves tracking physiological responses to improve the user experience, particularly regarding emotion and stress levels. So the key metrics involve eye movement, heart rate, GSR, and facial expression. Using GSR sensors to measure user stress levels during virtual reality simulations.

So now there are variables that capture this GSR data. There are smartwatches that capture this GSR data and basically help you find out about your well-being, how well you slept, how you're feeling, and so on. So social metrics focus on evaluating user behavior in a collaborative environment. So the key metrics are collaboration efficiency, communication clarity, social presence, and engagement in the group. An example is evaluating collaboration efficiency in a project management tool used by the team for task coordination.

So, for example, using Slack and many others, you can try to fit this here. So let's get back to our use case, the Braille learning app, how to do the evolution here. So a Braille learning app designed for visually impaired students needs to be evaluated to ensure accessibility, effectiveness, and ease of use. So here, basically, the evolution methods used are usability testing, heuristic evaluation, and surveys. So in usability testing, observe visually impaired students using the app to navigate the learning interface. Heuristic evaluation primarily involves applying accessible heuristics like screen reader compatibility to ensure that the app meets the accessibility guidelines.

Surveys basically collect feedback from users about their experience, satisfaction, and difficulty. You can also basically conduct a survey before learning, I mean, what their expectations are, what they are looking for, and so on. It is an alignment; it means it's good; if it is not, then probably there is some gap that you need to fill. So the outcome is refining the app's features based on feedback, such as simplifying navigation and adding clearer audio instructions. In our case, the project wave, as you can see here, Aditya and Mania are interacting with probably the users, a simulated user in this case, who are blindfolded. They are basically asked to follow a think-aloud protocol, where users vocalize their thoughts while performing the task, providing insight into their decision-making process.

Users of a language-learning app are asked to verbalize these steps as they attempt to complete a lesson. A set of users was instructed to execute various tasks to check the accessibility affordance of the interfaces. That is something we have done. And based on

that, again, you can collect the data; you can probably see if it meets the expectations you have for the users, so users can complete the task smoothly with an intuitive interface that meets their needs. One user faced initial confusion about successfully completing the task after the brief, and this is the average response time, detection time, accuracy, and so on, so you can collect all those required metrics.

Evaluation numbers based on the problem statement that you have. Similarly, you can perform remote usability testing, where users perform tasks from their own environment while researchers observe and record the interaction remotely. Similarly, our telemedicine app is tested remotely with doctors and patients to evaluate the app's usability. In the case of A-B testing, you can compare two versions of a design to see which is performing better in terms of user behavior and outcomes. Testing two different landing pages designed for an e-commerce website to measure the conversion rates is necessary.

So, as you can see, we have done the evaluation in low-fi fidelity. So if you recall from the previous lectures, we discussed the different lo-fi for this hand-drawn model. So this is one of them, and the second one is this. So you can probably find which one is best after evaluation. So, in this case, this one was probably the winner, chosen among these two options. and characteristics probably you can check button vibrator to gently nudge user buzzer to provide input output audio feedback and offers comfort and continuous tactile engagement and so on so again based on the different the requirement you can evaluate them and choose the one fit in your case similarly recruiting the what are the challenges in evolution One of the major challenges is recruiting the right users.

Because if you don't recruit the right user, it is probably being used by users who are not going to use it in real life. And in a way that can, again, there is no use for such a kind of evaluation. So it can be challenging to find users who represent the target demographic, especially for new applications. So, resource constraints and evaluation can be resource-intensive in terms of time, money, and manpower, definitely.

So, balancing between feedback and innovation. So too much focus on evaluation feedback can hinder creativity and innovation. And as a designer may become overly focused on user solutions, that is also something you don't want to be happening. So there should be some trade-off that you'll have to achieve, such as designing realistic test scenarios and creating scenarios that accurately reflect how users will interact with the product in real life, which is essentially very difficult. Similarly, evaluators exert different levels of control in lab and natural settings, and in crowdsourcing evaluation studies. So the results vary in different settings.

So data collection collected data to evaluate user experience goals such as challenge and

engagement. In addition to crowdsourcing quality control, we have also discussed in some of the earlier lectures how to control the quality of the collected data, especially in crowdsourcing scenarios. So a large number of participants can be recruited using Mechanical Turk and run experiments on the internet that are quick and inexpensive at the same time. To control the quality, we also need to remember the participants' rights and obtain their consent. That's where we discussed the ethical considerations of data collection, and we had a panel discussion on IRB as well. Participants need to be informed about why the evaluation is being done, what they will be asked to do, and their rights regarding informed consent.

Firms provided this information and acted as a contract between participants and the researchers. The design of the informed consent form, the evaluation process, data analysis, and data storage methods are typically approved by a high authority such as the IRB, as we discussed earlier. So things to consider when interpreting the data, reliability, so does the meta produce the same result on separate occasions? otherwise it may be by chance that is something you don't want to rely all your decision based on this so validity does the method measures what it is intended to measure ecological validity so does the environment of the evaluation distort the results biases are there biases that distort the result similarly scope how generalizable are the result is To summarize, evaluation is a critical part of SCI to ensure that products meet user expectations, are usable, and solve real-world problems effectively. So different evaluation types, formative, summative, diagnostic, longitudinal, et cetera, that we discussed, and techniques like usability testing, heuristics, eye tracking, et cetera, we discussed. So these have different purpose at various stages of product development.

The choice of technique depends on the product stage, user needs, and resource available. So participants need to be made aware of their rights. It is important not to over-generize findings from an evaluation. Participants need to be made aware of their rights, and it is important not to marginalize funding from an evaluation.

So user feedback is the fuel that drives design improvement. That's what Jared said. So field studies are evaluation studies that are carried out in natural settings to discover how people interact with technology in the real world. So often you need to collect data in natural settings, so that's what we can do. And similarly, field studies are innovate, involve the deployment of prototypes of technologies in natural setting that may be referred as in the wild studies.

So these are the further points which summarizes the evaluation. You can have a look. and evaluation metrics are critical in assessing systems' usability, performance, and overall user experience. The choice of metrics depends on the product's stage of

development, users' goals, and objectives. A combination of quantitative and qualitative metrics gives a comprehensive picture of how well a product meets user needs. So further, we are going to have a tutorial on evaluation by our brilliant TA, Rithvik.

We are going to have a demo on Adobe Illustrator by Shubhi. We'll be also having a demo on process of design process using drawing artwork using Adobe Express. And there are some additional reading resources. And if you want to give any suggestions or probably suggest Adobe Express to probably build some new feature that you think is revolutionary, please contact them. these are the further resources that you can have a look for further readings with this i stop here thank you so much see you in the next class