Human Computer Interaction (In English)

Prof. Rajiv Ratn Shah

Department of Computer Science and Engineering
Institute IIT Madras

Evaluation Techniques

Hi everyone, I am Ritwik Bamba and I am the teaching assistant for the course on human computer interaction. For today's tutorial, we will cover evaluation. First, we are going to be looking at what we have covered so far. We have looked at good design versus bad design, what makes a design good. inclusivity and accessibility and the design principles and how inclusivity, accessibility and usability come together to form an inclusive design. Next, we looked at the four basic activities in the interaction design process, which included establishing requirements and identifying needs, designing alternative designs, building interactive versions of the designs or prototyping, and then finally evaluating the designs and then further this design process goes on until we reach our final product right.

Next we covered the double diamond of interaction design, where we learned everything about our problem then narrow down to just one problem which was our problem statement. Then we considered every potential solution for our problem and then finally we find made the perfect solution which is our end product. Next, we covered user personas, how to create user personas, and empathy mapping, followed by a mental model, conceptual model, and the information architecture, which is a part of the conceptual model. For the user persona right here, we have the user's picture, their name, demographic information, what they do, their core needs, pain points, as well as the motivation.

It also includes a small quote that tells us about the personality of the user. Then we looked at what is cognition, cognitive load, what increases cognitive load, what decreases cognitive load. Then we move forward to prototyping, low fidelity prototypes, medium fidelity prototypes, and high fidelity prototypes. And then finally, we discussed the differences between low-fi, mid-fi, and high-fi. Now, moving on to the part that we are here for evaluation in context of human computer interaction or evaluation as we discussed during the interaction design process.

Well, evaluation simply is the testing the usability and functionality of the system. It occurs in collaboration with the users. It includes evaluating both the design as well as its implementation. It should be considered at all stages in the design process, which simply

means that we should be considering evaluation during all the stages of the process. This includes evaluation during establishing requirements, creating alternative designs, developing prototypes.

As you can see here, the process of creation, testing, evaluation goes on and on until we reach our final product. What are the goals of evaluation? Well, now we know what is evaluation and now we need to look at why evaluation. Well, first we need to assess the extent of the system functionality. Secondly, we have to assess the effect of interface on the user. And thirdly, we aim to identify specific problems which we may identify and then resolve them.

thus creating a better overall user experience. Well, evaluating implementations. For evaluating an implementation, we require artifacts. They may be simulation, prototype, or the full-scale implementation. By simulation, we mean that we simulate the conditions which allows the users to evaluate how the product is going to be functioning under certain circumstances.

It also means that the user gets to experience the real world application of the product. Secondly, we can also do the same using a prototype where we actually end up using the product as usability testing. Finally, we can also do the evaluation part on the full implementation, that is, the whole coded application, or as you would have known, some beta applications are also released to users. Those are examples of evaluations on full-scale implementations. We need to decide what type of data do we collect.

Well, for evaluation, we need data. And data, what kind of data? Well, data may be qualitative and quantitative. So first, we have look at qualitative data. It is the data that describes just the qualities or the characteristics. It is non-numeric and often is collected through direct observations or interviews.

It is usually more detailed. Well, how to collect quantitative data? We can have user interviews, focus groups, think-aloud protocols, and open-ended surveys, which allows the users to explain their experience in a much greater detail. What is the purpose? Well, we need to understand user experiences in detail and their motivations as well as their opinions. Secondly, we have quantitative data. Well, it is quite the opposite of qualitative data that it is a numerical data which can be measured and analyzed statistically.

Analyze statistically is the important part, right? How do we collect quantitative data? Well, we can see task completion times, error rates, and close ended service. Well, task completion time, think about it. If you are given a task, suppose a test. A test in school, you may need some time to complete it. And imagine if you are being graded on how

quickly you complete the test.

It is the best example of task completion time. Secondly, we have error rates. Well, simply, how much you score is the error rates. Finally, we have close-ended surveys, which aim to identify potential issues with the products and helps the designers improve upon them. What is the purpose? Well, we aim to measure performance, efficiency, and the accuracy of user interactions that happened during the testing phase.

Now we know what kind of data to collect. Now we need to decide whom to collect it from. So that we come to the next point, which is choosing participants. Well, the first and the foremost thing that we need to be careful about is that we need to ensure that the participants represent the actual users of the system and the representation isn't biased. For usability testing, 5 to 10 participants can reveal most usability issues.

Larger sample sizes may be needed for statistically significant findings in quantitative studies. We may need a bigger data set when we need to analyze the statistics of the data. Well, we also need to ensure that homogeneous groups are being used for testing. It is useful identifying small differences in control environments. At the same time, we also need to consider diverse groups which will provide broader insights, but then again, as we get broader insights we also get more variability in the results.

Finally, while using participants we need to consider what incentives do we need to offer they may be in form of compensation or money vouchers which in turn help encouraging their participation. We have to have the participants which represent the target groups and they should also have job specific vocabulary, knowledge or experiences. Think about it. If it's a system that is intended to make the lives of doctors better, won't it be better if we actually get doctors or medical students or nurses as participants? Once again, If your system is intended for engineers, won't it be great if we get engineering students to give us feedback for the same? Finally, we use incentives to encourage participation. You may give them goodies and or refreshments, encouraging more and more participation.

Now we look at experimental evaluation. What is experimental evaluation? It is simply the control evaluation of specific aspects of interactive behavior. The evaluator has a hypothesis that needs to be tested and the number of experimental conditions are considered is differ only in the value of the control variable. This is really important to note here. And well, experimental evaluation focuses on measuring how changes in the system impact user performance and experience.

Again, we come to the point, why do usability testing at all? Well, we can't tell how good a UI is until people use it or the users for which it is intended to be used. They may

actually use it. Only then can we find that how good the UI is. Well, experts, one may also think that why can't we just use experts instead of usability testing? Well, they may know too much or too less. Too much maybe that they may be going too much into the stuff where the solutions get over-engineered and thus make it difficult or they may actually not know enough about the real world problems that the users may actually tackle while testing.

And finally, it is hard to predict what the real users will actually do. We may not know what the users will do until we actually see for ourselves. Once we need to start with the usability testing, we need to create a usability testing proposal which contains the objective, the description of the system which is being tested, and some of the features that we are going to be testing, the task environments and the materials, how the testing is going to proceed, some details of the participants, Well, the methodology of how the usability testing is going to take place is one of the most important factors here. Then details of the tasks that the users are going to undertake. And finally, the test measures, how we are going to be evaluating the users, the parameters.

We need to get it approved and then reuse for the final report. It may seem tedious, but well, it helps you debug your test and identify and solve any problems that may have crept in okay so the next thing that we look at is selecting the tasks well tasks from a lo-fi design may actually be used the tasks may be shortened if they take too long in normal circumstances and they may also be shortened if the user requires background they actually don't have one more thing that we also need to take care of while selecting the tasks, that we don't put unnecessary or tasks that are out of scope for a user. We need to keep the tasks as relevant as possible and that the user's actions actually correspond to how the product will function in the real world. We also need to avoid bending tasks in direction of what our design supports best. We need to keep the usability testing free of any biases that may creep in.

And then we also need to be careful about tasks that are too fragmented. Well, because it simply does not represent a complete goal someone would be trying to accomplish with the application. Then we also need to include some of the experimental details. Well, the order of the tasks is really important here. We need to keep a simple order that the user start with simple tasks and move on to more complex tasks.

Well, unless they're doing in orders, we need to counterbalance the tasks. We also need to take care about the training part that how the tasks are going to be looked at. That it depends on how the real system will be used. So what we are trying to accomplish here is that we keep the tasks as relevant as possible and as close to the real usage that is going to happen with the solution. Well, we also need to take care of some of the external

factors like what if someone doesn't finish because they were allotted too much time.

Well, we need to look at these factors as well while we are working around with the details of the experiments. We also can do a pilot study. which helps you fix problems with the study. We can do two of the studies, which may start with the colleagues and then with the real users. This may actually help you improve the real study with the real users once you are done working with your colleagues.

You also need to look at how we compare two alternatives. The two alternatives are between the groups experiment and within the groups experiment. Different design solutions may be suited for different types of experiments. For the between groups experiment we actually take two users or two groups of test users and each of the systems uses only one of the systems and each of the group gives separate feedback for the separate system that they used. No fatigue is caused to the users because they only complete one condition.

For the within-groups experiment, we only have one group of test users, and each person uses both of the systems. And well, it requires a smaller sample size because each participant can provide data for both or all conditions. The benefit with within-groups experiments is that it is cheaper since it requires much less manpower. We also need to take care of some of the ethical considerations while we are evaluating the designs. First before the surveys or interviews we make the evaluation process completely voluntary with informed consent.

We also need to avoid any kind of pressurization on the participants that they need to take part in the studies. It should be voluntary. You also need to let them know that they can stop at any time at their own convenience. Next, we also stress that we are testing the system and not them. This is one of the most important parts here.

Finally, we make the collected data as much anonymous as possible. We may need to do some research, but that doesn't necessarily include using all of the data collected. Next, we move on to how do we actually measure the user preferences, how much the user likes or dislikes the systems. We can ask them to give you a rating on a scale of 1 to 10. or have them choose among the statements like the best UI I have ever used, or better than average, or worst UI I've ever used.

Well, it is kind of also hard to be sure about what the data will mean. Well, UIs are kind of subjective. So while users may like it, it may go not so well real world as well. If many give you low ratings, well, that may seem like a trouble for you because you may need to design the whole product once and for all. We also get some useful data by asking

what the users like, what they dislike, where they had trouble, where they had fun, the best part and the worst part.

This helps us identify the core problems and the core ideas that the users liked. And redundant questions are totally okay to make sure that the users completely understand and answer the questions properly. Well, now that we have collected the data, we need to interpret the test results. from the data to the actual interpretation part. First of all, we identify the patterns and trends, and we look for significant differences between the conditions.

And we also look out for recurring issues. We can use both types of data to explain or validate findings. We can use qualitative data and quantitative data together. That helps us in gathering much more insightful details about our designs. For instance, faster task times may actually correlate with positive feedback on ease of use. While we started off with a hypothesis, the end result of our evaluation is either supporting or refuting the hypothesis.

We need to determine if the results align with your initial hypothesis or they go completely against it. Finally, we get some actionable insights They translate the findings into design improvements. For example, if one component is disliked by a lot of users, or some of the users have mentioned that the product's components may actually increase cognitive load, we may need to improve upon the same design. Now that we have interpreted the results, we need to use them. We summarize the data and make a list of all the critical incidents.

Critical incidents are the ones that give us most insightful details, and positive as well as negative. We also include differences back to the original data, like some exact quotes from the users. Well, we also try to judge why each difficulty occurred. What was the cause of each difficulty so that doesn't occur again? We need to look at what the data is trying to tell us. What story is the data making up? Does the UI work the way you thought it would? Did the users take approaches that you expected? Or is there something missing? This is all that you can identify from the data that you collect from the users.

Finally, we update the tasks and rethink the design. We rate the severity of the critical incidents and the ease of fixing the critical incidents. We fix both severe problems and make the easy fixes. We also analyze the numbers that we collected in quantitative data collection. For example, suppose we get the task times as follows, the task completion time. While the mean may actually be 30, which is the average, the median is actually just 17.

5. Well, these discrepancies may actually account to different factors. These may include a small number of test results. Well, because the number of users were just 6, so these statistical measures may not be so much accurate. Then the results are very variable, once again as you can see in the data the standard deviation is 32, standard deviation is simply the dispersal from the mean. Well, this is what basic statistics can be used for, but statistics are well used and well suited for quantitative data with a much higher number of entries.

So now we will look at reporting the results. Well, we need to record what we did and what happened. So images and graphs will help people get it. And well, video clips can also be quite convincing. in telling people the story that we intend to tell, making up the data story and representing that in images, graphs, video clips, graphics, et cetera.

Next, we move on to an in-class activity. So imagine you're doing the testing of the usability of a food delivery app. Some of the metrics are collected during an evaluation we need to categorize each of these as qualitative or quantitative. Pause this and take a few minutes to do this yourself. Okay, now we are back to this.

The number of steps it takes for a user to place an order. It is definitely a quantitative measure because the measure would be clearly numeric, the number of steps. Second, we have the user comments about how easy it is to find specific cuisines. Well, explains comments about how easy it is. It is open-ended and hence qualitative. Next, we have the percentage of users who successfully complete the ordering process without any assistance.

It is definitely quantitative because it is the percentage of people or users, hence quantitative. Next, we have a participant saying the delivery tracking feature feels unreliable because it doesn't update in real time. Think about it. Well, it should be qualitative because it does not give out any numbers, numeric figures, hence qualitative. Next, we have the average rating users give to an app's search functionality on a scale of 1 to 5.

Definitely quantitative. Now we have some more questions. Please try doing them yourself. So we have the user describing their experience with the app as stressful during peak dinner hours. Well, it is definitely qualitative. The amount of time users spend browsing the menu before adding items to the cart.

Amount of time. Definitely quantitative. Feedback from a participant saying, I wish the app had a filter for dietary preferences like vegan or gluten-free. A feedback. A qualitative feedback. the number of times the user taps the back button while navigating

the app.

Definitely quantitative, the number of times. And finally, we have the percentage of participants who abandon their order after reaching the checkout screen. Well, it should also definitely be quantitative. Yeah. Now, these are some of the few resources that you can refer to for further readings about this topic on evaluation. Thank you so much.