User interface design

Week13: User interface Evaluation: The why, what, where, and when of evaluation, types of evaluation, usability Testing, Inspections

Nagulama Moses

Lecturer

Department Of Information Technology

Kumi University

Email: mnagulama@gmail.com

Outline

- Intended learning outcomes
- Introduction
- The Why, What, Where, and When of Evaluation
- Types of Evaluation
- usability Testing
- Inspections

Intended learning outcomes

- Explain why user interface (UI) evaluation is crucial in the development of digital products
- Define usability testing and its role in UI evaluation
- Describe inspection methods such as heuristic evaluation and cognitive walkthroughs

Introduction

- Imagine that you designed an app for young people to share music, gossip, and photos.
- You prototyped your first design and implemented the core functionality.
- How would you find out whether it would appeal to them and whether they will use it?
- You would need to evaluate it, but how?

- The main types of evaluation and the methods that you can use to evaluate design prototypes and design concepts at different stages in the lifecycle will be discussed.
- Evaluation is integral to the design process. It involves collecting and analyzing data about users' experiences when interacting with a sketch, prototype, or component of a system.
- Evaluation can happen during design, before a product is released, or even after a product is launched with the aim of improving or addressing a pain point reported by a customer.
- A central goal of evaluation is to improve its design.

- Evaluation focuses on both the usability of the product (learnability) and on the users' experiences when interacting with it (satisfying, enjoyable, etc.).
- Devices such as smartphones, iPads, e-readers, and also mobile apps continue to stimulate awareness about interaction design and usability.
- Evaluation enables designers to check that their design is appropriate and acceptable for the people who will use it.

- Evaluations can occur in a range of places such as in labs, people's homes, outdoors, work settings, and remotely, using digital video conferencing systems like Zoom or Teams, or via distributed design and evaluation systems.
- Product evaluations, such as the ranking and commenting systems that retailers use to get feedback about their products, can also be thought of as a kind of evaluation.

- Evaluations used to focus primarily on observing participants and measuring their performance during usability testing, experiments, or in natural settings, increasingly referred to as in-the-wild studies or research in the wild, to evaluate the design or design concept.
- But evaluation has become much broader, encompassing a range of methods, some of which involve working with participants remotely via digital and other technology.
- Others do not concern participants directly, such as modeling users' behavior and analytics.

- Modeling users' behavior provides an approximation of what users might do when interacting with an interface; these models are often done as a quick way of assessing the potential of different interface configurations.
- Analytics provide a way of examining the performance of an already existing product, such as a website, so that it can be improved.
- The level of control on what is evaluated varies; sometimes there is none, and in others there is considerable control over which tasks are performed and the context, such as in experiments.

- The methods selected will depend on several factors including what the evaluators want to find out, the type of product, when in the design the evaluation occurs, and logistical constraints such as cost and time.

The Why, What, Where, and When of Evaluation

- Conducting evaluations involves understanding not only why evaluation is important but also what aspects to evaluate, where evaluation should take place, and when to evaluate.

Why Evaluate?

- User experience involves all aspects of the user's interaction with the product.
- Nowadays people expect much more than just a usable product they also look for a pleasing and engaging experience from more products.
- Simplicity and taste are valued so that the product is a joy to own and use.
- Privacy is also important, especially for apps that record personal, health, and financial data.
- From a business and marketing perspective, well-designed products sell.

- Hence, there are good reasons for companies to invest in evaluating the design of products and to assess how popular the product is in the marketplace.
- Evaluation data enables designers to focus on real problems and the needs of different groups of people and make informed decisions about the design, rather than on debating what each other likes or dislikes.
- It also enables problems to be fixed before the product goes to be improved during its use.

What to Evaluate

- What to evaluate ranges from low-tech prototypes to complete systems, from a particular screen function to the whole workflow, and from appealing design to privacy, safety, and security features.
- Developers of a new web browser may want to know whether people can find relevant items faster using it.
- Developers of an ambient display may be interested in whether it changes people's behavior.

- Game app developers will want to know how engaging and fun their games are compared with those of their competitors and how long people will play them.
- Government authorities may ask if an AI system for controlling traffic lights results in fewer accidents or if a website complies with the standards required for people with disabilities.
- Makers of a toy may ask whether 6-year-olds can manipulate the controls, whether they are engaged by its furry cover, and whether the toy is safe for children.

- A company that develops health trackers may want to know whether people from different age groups and living in different countries like the size, color, and shape of the device.
- A software company may want to assess market reaction to its new home page design.
- A developer of smartphone apps for promoting environmental sustainability in the home may want to know if their designs are enticing and whether people continue to use their app after a period of time.

- Different types of evaluations will be needed depending on the type of product, the prototype or design concept, and the value of the evaluation to the designers, developers, and users.
- In the end, the main criteria are whether the design does what people need and want it to do and whether it is satisfying to use; that is, will they use it?

Where to Evaluate

- Where evaluation takes place depends on what is being evaluated.
- Some characteristics, such as web accessibility, may be evaluated in a lab because it provides the control necessary to investigate systematically whether all of the requirements are met.
- This is also true for design choices, such as choosing the size and layout of keys for a small handheld device for playing games.

- Increasingly apps, websites, and social media platforms are evaluated remotely by users distributed across the Internet through the use of remote tracking.
- For example, websites can be evaluated with a group of people using remote data logging software, which can also be used to create analytics.
- Similarly, logistical issues or ethical concerns may mean conducting remote evaluations is more feasible than running lab-based evaluations.

- However, despite advances being made in how to conduct remote evaluations, the user experience aspects, such as whether children enjoy playing with a new toy and for how long before they get bored, are still evaluated more effectively in natural settings, which are referred to as in-the-wild studies.
- Unlike a lab study, seeing children play in a natural setting will reveal how engaged they are when interacting with the toy and whether they play with it with their siblings or parents.

- In a lab study, the children are given instructions on what to do, which can guide and constrain their interactions with the toy.
- Remote studies of online behavior can be conducted to evaluate the interactions of participants with technology in their normal context of use.
- For example, in their own homes, places of work, or other settings where the researcher is not present, using logging software, video conferencing, participants' own video recording, or self-reflection tools, such as diaries.

When to Evaluate

- The stage in the lifecycle when evaluation takes place depends on the type of product and the development process being followed.
- For example, the product being developed could be a new concept, or it could be an upgrade to an existing product.
- It could also be a product in a rapidly changing market that needs to be evaluated to see how well the design meets current and predicted market needs.

- If the product is new, then considerable time may be invested in market research and discovering requirements.
- Once these requirements have been established, they are used to create initial sketches, a storyboard, or a prototype of the design ideas.
- These are then evaluated to see whether the designers have interpreted the requirements correctly and embodied them in their designs appropriately.

- The designs will be modified according to the evaluation feedback and new prototypes developed and subsequently evaluated.
- Evaluations conducted during design to check that a product continues to meet peoples' needs, are known as formative evaluations.
- Formative evaluations cover a broad range of design processes, from the development of early sketches and prototypes to tweaking and then perfecting a nearly finished design.

- Evaluations that are carried out to assess the success of a finished product are known as summative evaluations.
- If the product is being upgraded, then the evaluation may not focus on discovering new requirements but may instead evaluate the existing product to ascertain what needs improving.
- At other times, attention is focused on improving specific aspects, such as enhanced navigation or making the product more aesthetically pleasing.

Types of Evaluation

- We classify evaluations into three broad categories, depending on the setting, participants' involvement, and level of control. These are as follows:
- Controlled settings directly involving participants (labs): Participants' activities are controlled to test hypotheses and measure or observe certain behaviors. The main methods are usability testing and experiments.

- Natural settings involving people (examples are online communities and products that are used in public places): There is little or no control of peoples' activities to determine how the product would be used in the real world. The main method used is in-the-wild studies.
- Any settings not directly involving participants: Consultants and researchers critique, predict, and model aspects of participants' interactions with the product to identify the most obvious usability problems.
- The range of methods includes inspections, heuristics, walkthroughs, models, and analytics.

- There are pros and cons of each evaluation category. For example, lab-based studies are good at revealing usability problems, but they are poor at capturing context of use;
- In the- wild studies are good at demonstrating how people use technologies in their intended setting, but they are often time-consuming and more difficult to conduct and modeling and predicting approaches are relatively quick to perform, but they can miss unpredictable usability problems and subtle aspects of the user experience.

- Similarly, analytics are good for tracking the use of a website but are not good for finding out how users feel about a new color scheme or why they behave as they do.
- Deciding on which evaluation approach to use is determined by the goals of the project and on how much control is needed to find out whether an interface or device meets those goals.
- This includes finding out how people use it, whether they like it, and what problems they experience with the functions.

- In turn, this requires determining how they carry out various tasks using the interface operations.
- A degree of control is needed when designing the evaluation study to ensure participants try all of the tasks and operations for which the app or system is designed.

Controlled Settings Involving Participants

- Experiments and user tests are designed to control what participants do, when they do it, and for how long.
- They are designed to reduce outside influences and distractions that might affect the results, such as people talking or music playing in the background.
- The approach has been extensively and successfully used for many years to evaluate software applications running on laptops and other devices.

❖**Usability Testing**

❖This approach to evaluating user interfaces involves collecting data using a combination of methods in a controlled setting, for example, experiments that follow basic experimental design, observation, interviews, and questionnaires.

❖Often, usability testing is conducted in labs, although increasingly interviews and other forms of data collection are being conducted remotely via phone and digital communication (for instance, through Teams or Zoom) or in natural settings.

- The primary goal is to determine whether an app or system is usable for the tasks conducted by the people for whom it was designed. This involves investigating how typical users perform on typical tasks.
- By typical, we mean the people for whom the system is designed for and the activities that it is designed for them to be able to do.
- It often involves comparing the number and kinds of errors that are made between versions and recording the time that it takes the people using it to complete a task.

- As participants perform the tasks, they may be recorded on video, and their interactions may also be recorded by logging software.
- User satisfaction questionnaires and interviews can also be used to elicit opinions about how they liked the experience of using the system.
- This data can be supplemented by observation at product sites to collect evidence about how the product is being used in the workplace or in other environments.

- Observing people’s reactions to an interactive product has helped developers reach an understanding of usability issues, which would be difficult for them to glean simply by reading reports or listening to presentations.
- The qualitative and quantitative data that is collected using these different techniques are used in conjunction with each other to form conclusions about how well a product meets the needs of its users.

In-the-Wild Studies

- In-the-wild studies are a way of observing technology usage in natural settings with little or no evaluator involvement or presence.
- In the early days of usability evaluation, evaluators realized that they needed to understand how prototypes and other design products were used in natural settings outside of the lab by observing and talking with users.
- There has been a trend toward conducting in-the-wild studies in which evaluators had less and less control over and direct involvement with the study participants.

- The goal of in-the-wild studies is to evaluate products with people in their natural settings, primarily to:
- Help identify settings for new technologies
- Inform the requirements for a new technology design based on in situ observations
- Introduce new technology interventions or inform deployment of existing technology in new contexts
- Enable extended engagement with the technology

- Doing in-the-wild studies to evaluate a product can provide more ecological validity as they assess the fitness of a product for the intended audience in an everyday context while enabling unforeseen issues related to the environment of use to be revealed.
- In-the-wild studies also enable usability measures (e.g., time spent, features explored) to be recorded in a real-world setting, such as how social media is used by target audiences on their smartphones over a period of time.
- Methods that are typically used for in-the-wild studies are observation, interviews, and interaction logging

- The data takes the form of events and conversations that are typically recorded through audio or video recording, or by the participants as diaries and notes.
- The goal of the evaluators is to be unremarkable and not to affect what people do during the evaluation.
- However, it is inevitable that some methods will influence how people behave. For example, the use of cameras can make people feel self-conscious and change how they behave or choose to interact with the technology intervention.

- In-the-wild studies involve looking at how new technologies or prototypes are deployed and used by people in various natural settings, such as outdoors, in public places, and in their homes.
- Sometimes, a prototype that is deployed is called a disruptive technology, because the aim is to discover how it displaces an existing technology or practice.
- In moving into the wild, evaluators inevitably give up control over what is being evaluated in order to observe how people approach and use or don’t use technologies in their everyday lives.

- For example, an evaluator might be interested in observing how a new mobile navigation device will be used in urban environments.
- To conduct an in-the-wild study where a specific device is being introduced, researchers need to recruit people who are willing to use the device for a few weeks or months in their natural surroundings.
- They might then tell the participants what they can do with the device. Other than that, it is up to the participants to decide how to use it and when, as they move among work or school, home, and other places.

Settings Not Involving Participants

- Evaluations that take place without involving participants are conducted in settings where the researcher has to imagine or model how an interface is likely to be used.
- Inspection methods are commonly employed to predict user behavior and to identify usability problems based on knowledge of usability, users' behavior, the contexts in which the system will be used, and the kinds of activities that people undertake.

- Examples include heuristic evaluation that applies knowledge of typical users guided by rules of thumb and walk-throughs that involve stepping through a scenario or answering a set of questions for a detailed prototype.
- There are tailored heuristics for evaluating most product types, including web-based products, mobile apps, collaborative technologies, conversational agents, computerized toys, games, information visualizations, and more.
- One of the problems with using heuristics is that designers can sometimes be led astray by finding that heuristic evaluation is not as accurate as it appeared to be at first.

- This problem can arise from different sources, such as a lack of experience and the biases of UX researchers who conduct the heuristic evaluations.
- Cognitive walk-throughs involve simulating a user's problem-solving process at each step in the human-computer dialogue and checking to see how users progress from step to step in these interactions.
- During the last 20 years, cognitive walk-throughs have been used to evaluate smartphones, large displays, and other applications, such as public displays and interface learnability.

- A key feature of cognitive walk-throughs is that they focus on evaluating designs for ease of learning.
- Analytics is a technique for logging and analyzing data either at a customer's site or remotely.
- Web analytics is the measurement, collection, analysis, and reporting of Internet data to understand and optimize web usage.

- Examples of web analytics include the number of visitors to a website home page over a particular time period, the average time users spend on the home page, which other pages they visit, or whether they leave after visiting the home page.
- For example, Google provides a commonly used approach for collecting analytics data that is particularly useful for evaluating design features of a website

Selecting and Combining Methods

- The three broad categories identified earlier provide a general framework to guide the selection of evaluation methods.
- Often, combinations of methods are used across the categories to obtain a richer understanding.
- For example, sometimes usability testing conducted in labs is combined with observations in natural settings to identify the range of usability problems and find out how users typically use a product.
- There are both pros and cons for controlled and uncontrolled settings.

- The benefits of controlled settings include being able to test hypotheses about specific features of the interface where the results can be generalized to the wider population.
- A benefit of uncontrolled settings is that unexpected data can be obtained that provides quite different insights into people's perceptions and their experiences of using, interacting, or communicating through the new technologies in the context of their everyday and working lives.

Opportunistic Evaluations

- Evaluations may be detailed, planned studies, or opportunistic. The latter explorations are generally done early in the design process to provide designers with feedback quickly about a design idea.
- Getting this kind of feedback is important because it confirms whether it is worth proceeding to develop an idea into a prototype.
- Typically, these early evaluations are informal and do not require many resources. For example, the designers may recruit a few local people and ask their opinions.

- Getting feedback this early in design provides feedback when it is easier to make changes to an evolving design.
- Opportunistic evaluations with users can also be conducted to improve the target audience so that subsequent evaluation studies can be more focused.
- Opportunistic evaluations can also be conducted in addition to more formal evaluations.

Usability Testing

- Traditionally the usability of products was tested in laboratories or other controlled settings.
- Initially, usability testing focused on desktop applications, such as websites, word processors, and search tools. Now it is also important to test the usability of apps and other digital products.
- Performing usability testing in a laboratory, or in a temporarily assigned controlled environment, enables designers to control what participants do and also to control the environmental and social influences that might impact the participants' performance.

- The goal is to test whether the product being developed is usable by the intended participants to achieve the tasks for which it was designed and whether participants are satisfied with their experience.
- For some products, such as games, designers also want to know whether their product is enjoyable and fun to use; this is an important business goal.
- Increasingly, aesthetic design has become important, especially for websites designed to sell products

Methods, Tasks, and Users

- Collecting data about people's performance on predefined tasks has always been a central component of usability testing.
- A combination of methods is used to collect data, which usually includes logged keystrokes, mouse and other movements, such as swiping, pointing, and dragging objects, and video recordings of participants, including their facial expressions.
- Sometimes, participants are asked to describe out loud what they are thinking and doing (the "think-aloud" technique) while carrying out tasks.

- In addition, participants complete a user satisfaction questionnaire about their reactions to using the product using rating scales and answering questions.
- Structured and semi structured interviews may also be used to collect information about what the participants liked and did not like about the product.
- Evaluators may also observe and collect data about how the product is used in natural settings such as offices, hotels, people's homes, etc.

- Examples of the tasks that are typically given to participants include searching for information, comparing different icons, navigating through different menus, and downloading apps.
- Performance times and the number and type of actions performed by participants are the two main performance measures.
- Obtaining these two measures involves recording the time it takes typical users to complete a task, such as finding a topic on a website, and the number of errors that users make, such as when selecting incorrect menu options or links.

- The following quantitative performance measures were identified and used as a baseline for collecting user performance data:
- Number of people completing a task successfully
- Time to complete a task
- Time to complete a task after a specified time away from the product
- Number and type of errors per task

- Number of errors per unit of time
- Number of times people navigate to an item such as online help
- Number of people making the same or similar errors
- A key concern when doing usability testing is the number of participants that should be involved:
- Early research suggests that 5 to 12 is an acceptable number, though more is often regarded as being better because the results represent a larger and often broader selection of the participant population.

Labs and Equipment

- Large companies, such as Microsoft, Google, and Apple, and some government agencies test their products in usability labs that consist of a main testing lab with recording equipment and an observation room where the evaluators can watch what is going on and how the data collected is being analyzed.
- There may also be a reception area where participants can wait, a storage area, and a viewing room where observers such as designers, product managers, and others can watch.

- These lab spaces can vary and can be arranged to mimic features of the real world superficially.
- For example, when testing an office product or a product for use in a hotel reception area, the lab can be set up to resemble those environments.
- Soundproofing and lack of windows, co-workers, and other workplace and social distractions are eliminated so that participants can concentrate on the tasks that have been set up for them to perform.

- While controlled environments like these enable researchers to capture data about participants' uninterrupted performance, the impact that real-world interruptions can have on usability is not captured.
- Furthermore, purpose-built labs are expensive for small organizations to maintain.
- Typically, large company labs contain two to three wall-mounted video cameras that record the participants' behavior, such as hand movements, facial expressions, and general body language.
- Microphones are placed near where the participants will be sitting to record their comments.

Inspections: Heuristic Evaluation and Walk-Throughs

- Sometimes, it is not practical to involve participants in an evaluation because they are not available, there is insufficient time, or it is difficult to find people.
- In such circumstances, other people, often referred to as experts or researchers, can provide feedback. Such people should be knowledgeable about both interaction design and the needs and typical behavior of users.

- Various inspection methods were developed as alternatives to usability testing in the early 1990s, drawing on software engineering practice where code and other types of inspections are commonly used.
- Inspection methods for interaction design include heuristic evaluations and walk-throughs, in which researchers examine the interface of an interactive product,and suggest problems that people would likely have when interacting with the product.
- One of the attractions of these methods is that they can be used at any stage of a design project.
- They can also be used to complement other types of usability testing.

Heuristic Evaluation

- In heuristic evaluation, evaluators, guided by a set of usability principles known as heuristics, evaluate whether user-interface elements, such as menus, navigation structure, sound, online help, conversational agents, and so on, conform to tried-and-tested principles.
- These heuristics closely resemble high-level design principles (such as making designs consistent, reducing memory load, and using terms that users understand).
- In addition, many researchers and practitioners have converted design guidelines into heuristics that are then applied in heuristic evaluation.

- The original set of heuristics evaluation include the following:
- Visibility of System Status: The system should always keep users informed about what is going on, through appropriate feedback and within a reasonable time.
- Match Between System and the Real World: The system should speak the users' language, with words, phrases, and concepts familiar to the user, rather than system-oriented terms.
- User Control and Freedom: Users often choose system functions by mistake and will need a clearly marked emergency exit to leave the unwanted state without having to go through an extended dialog.

- ❖Consistency and Standards: Users should not have to wonder whether different words, situations, or actions mean the same thing. The system should follow platform conventions.
- ❖Error Prevention: Rather than just good error messages, the system should incorporate careful design that prevents a problem from occurring in the first place.
- ❖Recognition Rather Than Recall: Minimize the user's memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the dialog to another.

- Flexibility and Efficiency of Use: Accelerators unseen by the novice user may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users.
- Aesthetic and Minimalist Design: Dialogs should not contain information that is irrelevant or rarely needed. Every extra unit of information in a dialog competes with the relevant units of information and diminishes their relative visibility.

- Help Users Recognize, Diagnose, and Recover from Errors: Error messages should be expressed in plain language (not codes), precisely indicate the problem, and constructively suggest a solution.
- Help and Documentation: Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, list concrete steps to be carried out, and not be too large.

Walk-Throughs

- Walk-throughs offer an alternative approach to heuristic evaluation for predicting users' problems without doing user testing.
- As the name suggests, walk-throughs involve walking through a task with the product and noting problematic usability features.
- While most walkthrough methods do not involve users, others, such as pluralistic walk-throughs, involve a team that may include users as participants, as well as developers and usability specialists.

Cognitive Walk-Throughs

- A cognitive walkthrough (CTW) is a usability inspection method that uses a cognitive model of human performance to assess the usability of a product or system.
- CTWs are typically conducted by a group of experts who walk through a set of tasks that a user would typically perform with the product or system.

- At each step of the task, the experts ask themselves a series of questions about the user's goals, knowledge, and actions.
- The goal of a CTW is to identify any potential usability problems with the product or system, so that they can be addressed before the product or system is released to users.

Steps involved in cognitive walk-throughs

1. The characteristics of typical users are identified and documented, and sample tasks are developed that focus on the aspects of the design to be evaluated.

❖A description, mock-up, or prototype of the app or system to be developed is also produced, along with a clear sequence of the actions needed for people to complete the task.

2. A designer and one or more UX researchers come together to do the analysis.

3. The UX researchers walk through the action sequences for each task, placing it within the context of a typical scenario. As they do this, they try to answer the following questions:

a) Will the correct action be sufficiently evident to the user?

b) Will the user notice that the correct action is available?

c) Will the user associate and interpret the response from the action correctly?

❖In other words, will users know what to do, see how to do it, and understand from feedback whether the action was completed correctly or not?

4. As the walk-through is being done, a record of critical information is compiled.

a) The assumptions about what would cause problems and why are identified.

b) Notes about side issues and design changes are made.

c) A summary of the results is compiled.

❖5. The design is then revised to fix the problems presented. Before making the fixes, insights derived from the walk-through are often checked by testing them with real users.

- When doing a cognitive walk-through, it is important to document the process, keeping account of what works and what doesn't.
- A standardized feedback form can be used in which answers are recorded to each question.
- Any negative answers are carefully documented on a separate form, along with details of the product, its version number, and the date of the evaluation.
- It is also useful to document the severity of the problems. For example, this includes how likely a problem is to occur and how serious it will be for users.

Pluralistic Walk-Throughs

- A pluralistic walk-through (PWT) is a usability inspection method that is similar to a cognitive walkthrough, but it involves a more diverse group of participants, including users, developers, and other stakeholders.
- The goal of a PWT is to identify usability problems from a variety of perspectives, and to develop solutions that are acceptable to all stakeholders.

Steps for pluralistic walk-through

- Identify the tasks to be evaluated. This can be done by interviewing users or by observing them performing tasks.
- Recruit a group of participants. The group should include users, developers, and other stakeholders.
- Prepare the environment. This may involve setting up a prototype of the product or system, or using screenshots or videos.

- Walk through the tasks one by one. For each task, the facilitator should ask the participants the following questions:
  - What is the user trying to do?
  - What information does the user need to complete the task?
  - Are the controls easy to find and use?
  - Is the feedback clear and helpful?
- Identify usability problems and discuss solutions. The participants should discuss any usability problems that they identify, and try to develop solutions that are acceptable to all stakeholders.

References

- Preece, R., Rogers, Y., & Sharp, H. (2023). Interaction Design: Beyond Human-Computer Interaction. J. Wiley & Sons.
- Norman, D. A. (2013). The Design of Everyday Things: Revised and Expanded Edition. Basic Books.

Thank you

Next Lecture We Shall Look At

Lab practicals