HCI 2022

Evaluation: Usability Testing

Human Computer Interaction

Luigi De Russis, Alberto Monge Roffarello

Academic Year 2022/2023

Evaluation Goal (recap)

- «Evaluation tests the usability, functionality, and acceptability of an interactive system»
  - According to the design stage (sketch, prototype, … final)
  - According to the initial goals
  - Alongside different dimensions
  - Using a range of different techniques
- Very wide (and a little bit vague) definition
- The idea is to identify and correct problems as soon as possible

Evaluation Approaches (recap)

- Evaluation may take place:
  - In the laboratory
  - In the field
- Involving users:
  - Experimental methods
  - Observational methods
  - Query methods
  - Formal or semi-formal or informal
- Based on expert evaluation:
  - Analytic methods
  - Review methods
  - Model-based methods
  - Heuristics
- Automated:
  - Simulation and software measures
  - Formal evaluation with models and formulas
  - Especially for low-level issues

Experts vs. Users

Expert methods,

e.g., heuristic evaluation

- Faster (1-2h per evaluator)
- Results are pre-interpreted (thanks to the evaluators)
- Could generate *false positives*
- Might *miss* some problems
- **useful for filtering and refining the design!**

User studies,

e.g., usability testing

- Need to develop software, and prepare the set-up
- More accurate (by definition!)
  - Actual users and tasks
- **tends to occur in the later stages of development**

Lab vs. Field

Evaluation in Lab

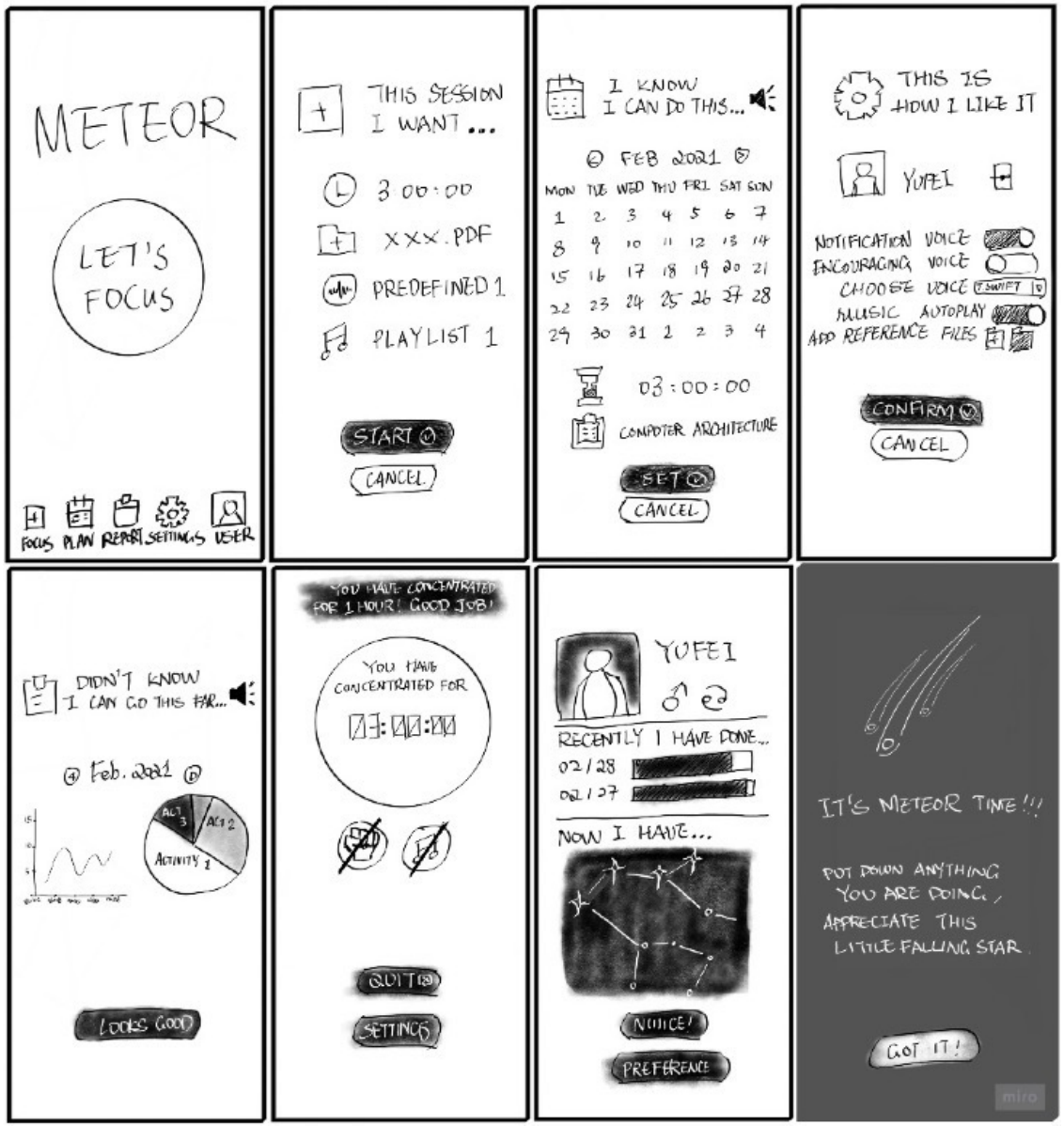
- Advantages
  - specialist equipment available
  - uninterrupted environment
- Disadvantages
  - lack of context
  - difficult to observe several users cooperating
- Appropriate
  - if system location is dangerous or impractical
  - for constrained single user systems to allow controlled manipulation of use

Evaluation in the Field

- Advantages
  - natural environment
  - context retained (although observation may alter it)
  - longitudinal studies possible
- Disadvantages
  - distractions
  - noise
- Appropriate
  - where context is crucial
  - for longitudinal studies

Example: Lab vs. Field

- Through the Meteor app, *"users can use straight-forward, effective reminders to efficiently finish their tasks."*
- Through **lab** and **field** studies I can test different aspects of the app and learn different things!



Involving Users: Experimental Methods

Usability/User Testing

- "Let's find someone to use our app, so that we will get some feedback on how to improve it."
- anecdotal, mostly
- observation-driven

Controlled Experiments

- "We want to verify if users of our app perform task X faster/…/with fewer errors than our competitor's app."
- scientific
- hypothesis-driven

Controlled Experiments

- Controlled evaluations of **specific** aspects of interactive behavior
  - typically in lab
- Provide empirical evidence to support a particular claim or hypothesis
- The evaluator chooses a hypothesis to test, which can be determined by measuring some attribute of participant behavior

Controlled Experiments

- Hypothesis
  - the prediction of the outcome of the study, what you would like to demonstrate
  - framed in terms of *variables*
  - in the form of a **null hypothesis**, to be disproved
- Variables
  - things to manipulate and measure, to test the hypothesis
- Subjects (participants)
  - representative, sufficient sample
  - sample size: at least double the number suggested by Nielsen for usability tests (5)
  - vital to the success of any experiment

Usability Testing

- The Nielsen-Norman Group defines usability testing in this way:

  In a usability-testing session, a researcher (called a "facilitator" or a "moderator") asks a participant to perform tasks, usually using one or more specific user interfaces. While the participant completes each task, the researcher observes the participant's behavior and listens for feedback.

- Sometimes called User Testing

- Typically, the goal is to:
  - Identify problems in the design
  - Uncovering opportunities to improve
  - Learning about the target user's behavior and preferences

Usability Testing

Core Elements of Usability Testing

Facilitator
Guides the participant through the test process

Tasks
Realistic activities that the participant might actually perform in real life

Participant
Realistic user of the product or service being studied

NNGROUP.COM NN/g

Usability Testing: Flow of Information

NNGROUP.COM NN/g

Facilitators and Observers

- The facilitator guides the participant through the test process
- It gives instructions, answers the participant's questions, and asks follow-up questions
- **It should not influence the participants' behavior!**
- There can be dedicated figures serving as **observers**:
  - They can focus on observing participants, e.g., by taking notes, without interacting with them

Tasks

- Tasks in a usability test are realistic activities that the participant might perform in real life:
  - *Your printer is showing "Error 5200". How can you get rid of the error message?*
  - *You're considering opening a new credit card with Wells Fargo. Please visit wellsfargo.com and decide which credit card you might want to open, if any.*
- Task instructions can be delivered to the participant verbally or written on a sheet of paper
- Ask participants to read the tasks out loud!

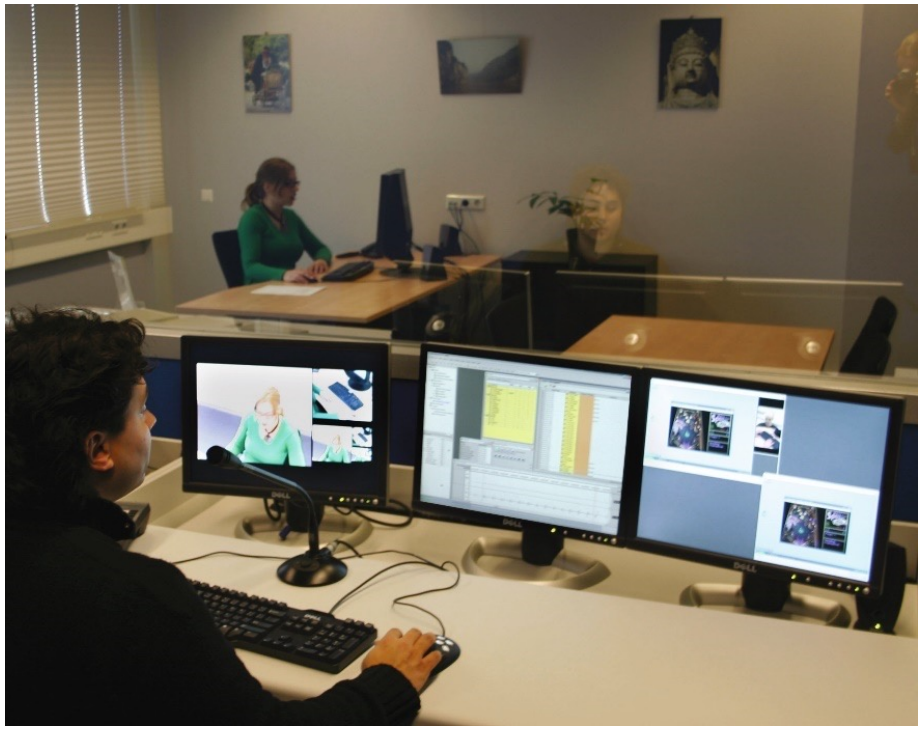
Participants

- Participants should be realistic users of the product or service being studied:
  - they already use the system
  - they are “similar” to the target user group
- Participants should represent the intended user communities, with attention to:
  - background in computing and experience with the task
  - motivation, education, and ability with the natural language used in the interface

Participants

- **How many** participants do you need?
  - 5!
  - https://www.nngroup.com/articles/how-many-test-users/
- Testing costs increase with each additional study participant, yet the number of findings quickly reaches the point of diminishing returns
- Rule of thumb: if you have a big budget, spend it on additional studies, not more users in each study!

Usability Testing Labs



- The usability lab usually consists of two areas
  - the testing room
  - the observation room
- The testing room is typically smaller and accommodates a small number of people
- The observation room can see into the testing room typically via a one-way mirror
  - it is larger and can hold the facilitators with ample room to bring in others, such as the developers of the product being tested

Type of Usability Testing

- Usability testing can be either **qualitative** or **quantitative**:
  - Qualitative usability testing focuses on collecting insights, findings, and anecdotes about how people use the product or service
  - Quantitative usability testing focuses on collecting metrics that describe the user experience, e.g., task success and time on task
- Usability testing can be either performed **in-person** or **remotely**

Usability Testing: 3 Steps

1. **Plan**
   - who are your participants? what are you going to test, where, and how?
2. **Run**
   - one participant at time, multiple sessions
   - collect data about the interactive system/interface
3. **Analyze**
   - extract information from the collected data, both qualitative and quantitative

Plan

Usability Testing

Usability Testing: Plan

- Choose **who** you will involve in the test
  - who are your (target) users?
- **How many** participants do you need?
  - 5 are typically enough!
- Decide who and which **roles** you are going to "play"
  - you need at least a facilitator of the session
  - other 1-2 people may serve as note-takers and observers
  - **N.B.** developers, designers, creators, … of the interactive system in evaluation must not serve as facilitators!

Usability Testing: Plan

- Choose **which task(s)** you are going to ask your participants to perform
  - tasks may be introduced with a scenario
  - they must be *concrete* and with a *clear goal*
  - between *5-10* tasks, typically
- Define detailed success/failure **criteria** for each task

Usability Testing: Plan

- Choose any **methodology** you are willing to apply
  - think-aloud, cooperative evaluation, …, and/or none
    - more details in a few slides
  - and for which tasks you are going to use it

Usability Testing: Plan

- Decide whether you need or want to ask any **additional information**
  - before and/or after the test
  - before and/or after each task
  - before and/or after a meaningful group of tasks

- Select which **equipment** you will need
  - also with respect to the criteria and methodology you define

- Prepare an **informed consent form** for participants to fill

Usability Testing: Plan

- Decide whether to have a **debriefing** session at the end of the test
  - for each participant
  - observers and note-takers can ask general and specific questions, to better understand some pathways or comments
- Develop a **written test protocol** ("script") for consistency among sessions
  - step-by step instructions with all the needed questions and forms
  - often down to the exact words that the facilitator will say
  - the appendix may contain a table with all tasks and their metrics
- *Practice* your script with friends or colleagues
  - to fix obvious bugs so that you do not waste (yours and users') time

Informed Consent Form

- Professional ethics practice is to ask all participants to read, understand, and sign a statement which says:
  - I have freely volunteered to participate in this experiment
  - I have been informed in advance what my task(s) will be and what procedures will be followed
  - I have been given the opportunity to ask questions and have had my questions answered to my satisfaction
  - I am aware that I have the right to withdraw consent and to discontinue participation at any time, without prejudice to my future treatment
  - My signature below may be taken as affirmation of all the above statements; it was given prior to my participation in this study

Tasks

- Define tasks according to the main goals that target users may have on your system
- Rather than simply ordering test users to "do X" with no explanation, it's better to situate the request within a short **scenario:**
  - it sets the stage for the action and explains why the user is "doing X."
- A task scenarios for usability testing need to provide **context** so users engage with the interface and pretend to perform it as if they were at home or in the office

Tasks

- Make the task **realistic**:
  - User goal: Browse product offerings and purchase an item.
  - Poor task: Purchase a pair of orange Nike running shoes.
  - Better task: Buy a pair of shoes for less than $40.
- Make the task **actionable:**
  - User goal: Find movie and show times.
  - Poor task: You want to see a movie Sunday afternoon. Go to www.fandango.com and tell me where you'd click next.
  - Better task: Use www.fandago.com to find a movie you'd be interested in seeing on Sunday afternoon.

Tasks

- Avoid giving **clues** and **describing the steps**
  - User goal: Look up grades.
  - Poor task: You want to see the results of your midterm exams. Go to the website, sign in, and tell me where you would click to get your transcript.
  - Better task: Look up the results of your midterm exams.

Metrics

- For success/failure criteria and additional information
- *Subjective* metrics, i.e., questions you ask participants:
  - prior to the session, e.g., background info
  - after each task scenario is completed, such as ease and satisfaction questions about the task
  - overall ease of use, satisfaction, and likelihood to use/recommend at the end
- *Quantitative* metrics
  - what you will be measuring in your test, e.g., successful completion rates, error rates, time on task

Metrics

Successful Task Completion	A task is successfully completed when the participant indicates they have found the answer or completed the task goal.	Boolean value, 0-100 scale, …
Critical Errors	Deviations at completion from the targets of the task, so that the participant cannot finish the task. Participant may or may not be aware that the task goal is incorrect or incomplete.	Absolute or relative number
Non-Critical Errors	Errors that are recovered by the participant and do not result in the participant's ability to successfully complete the task. These errors result in the task being completed less efficiently.	Absolute or relative number, or they may affect the "successful task completion"
Error-Free Rate	The percentage of participants who complete the task without any errors.	Relative number

Metrics

Time On Task	The amount of time it takes the participant to complete the task.	Time
Subjective Measures	Self-reported participant ratings for satisfaction, ease of use, ease of finding information, etc.	Likert Scale
Likes, Dislikes and Recommendations	What participants liked the most about the system, what they liked least, any recommendations for improving it, etc. Typically at the end of the session or a meaningful part of it.	Free text

Reliable and validated questionnaires exist for subjective measures and open questions

Measuring Success

- Task success or completion is one of the more common metrics used in user experience
- In its simples form, it's a binary metric
  - How do we account for cases of partial success?
- We can define several levels of success, depending on the evaluated system and tasks

Measuring Success: Example

- TASK: you have to book a room of a given size, in a given date, and for a certain amount of time by exploiting the Politecnico's website
- We might define several levels of success, e.g.,:
  - **complete success:** the user book the room with no error, exactly as specified
  - **success with one minor issue:** the user book the room but select a wrong size
  - **success with a major issue:** the user book the room but enters the wrong date or amount of time
  - **failure:** the user is not able to book the room

Measuring Success: Example

Level of success	Number of users (out of 100)	How you report it
Complete success	20	20% of our participants were able to complete the task successfully with no error. Based on this result, we expect that between 13% and 29% (*) of our general user population will complete the task with no error.
Success with a minor issue	35	35% of our participants placed an order but had a minor issue. Based on this result, we expect that between 26% and 45% (*) of our general user population will complete the task with a minor error.
Success with a major issue	30	30% of our participants placed an order but encountered a major issue. Based on this result, we expect that between 22% and 40% (*) of our general population will complete the task with a major error.
Failure	15	15% of our participants were not able to place the order. Based on this task, we expect that between 9% and 23% (*) of our general population will not be able to place an order.

(*) In this table, the ranges represent 95% confidence intervals calculated using the Adjusted Wald method.

Taken from https://www.nngroup.com/articles/success-rate-the-simplest-usability-metric/

Post-Task Questionnaire: SEQ

- *Single Ease Question (SEQ)*

- Post-task questionnaires need to be short (1–3 questions) to interfere as little as possible with the flow of using the system in a session
- SEQ exemplifies this concept in a useful and simple manner
  - experimentally validated
  - reliable, valid, and sensitive
- It asks the user to rate the difficulty of the activity they just completed, from Very Easy to Very Difficult on a 7-point Likert scale

Post-Test Questionnaire: SUS

- *System Usability Scale (SUS)*
  - a "quick and dirt" (but trustable) usability scale
  - invented by John Brooke in 1986
- It measures the **perceived usability** of a system
- A 10-item Likert-scale questionnaire
  - each question has 5 response options
- It produces a score from 0-100
  - not equivalent to a percentage score!
- A SUS score above 68 is considered **above average**


1. I think that I would like to use this system frequently.
1. Strongly Disagree 2. 3. 4. 5. Strongly Agree

2. I found the system unnecessarily complex.
1. Strongly Disagree 2. 3. 4. 5. Strongly Agree

3. I thought the system was easy to use.
1. Strongly Disagree 2. 3. 4. 5. Strongly Agree

4. I think that I would need the support of a technical person to be able to use this system.
1. Strongly Disagree 2. 3. 4. 5. Strongly Agree

5. I found the various functions in this system were well integrated.
1. Strongly Disagree 2. 3. 4. 5. Strongly Agree

6. I thought there was too much inconsistency in this system.
1. Strongly Disagree 2. 3. 4. 5. Strongly Agree

7. I would imagine that most people would learn to use this system very quickly.
1. Strongly Disagree 2. 3. 4. 5. Strongly Agree

8. I found the system very cumbersome to use.
1. Strongly Disagree 2. 3. 4. 5. Strongly Agree

9. I felt very confident using the system.
1. Strongly Disagree 2. 3. 4. 5. Strongly Agree

10. I needed to learn a lot of things before I could get going with this system.
1. Strongly Disagree 2. 3. 4. 5. Strongly Agree

SUS: Questions

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

SUS: Scoring

To **calculate** the SUS score of your system:

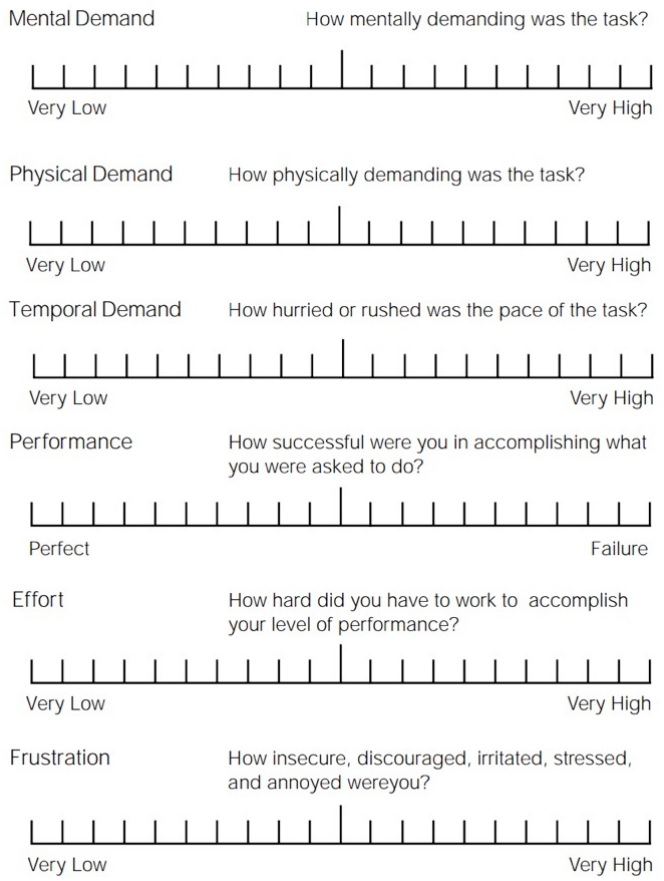
1. Each answer is 1-5 (X)
2. For every odd-numbered question, subtract 1 from the score (X-1)
   - e.g., the answer for question 1 is 4, so its score is 4-1 = 3
3. For every even-numbered question, subtract the score from 5 (5-X)
   - e.g., the answer for question 2 is 4, so its score is 5-4 = 1
4. Sum the scores from even and odd-numbered questions
5. Multiply the total by 2.5

SUS: Advantages and Disadvantages

- Advantages
  - Score reliability has been evaluated over the decades and it is on par with more complex and costly methods
  - Free, quick, and simple
  - Quite used in industry
  - Applicable to a wide range of technologies, systems, and products
- Disadvantages
  - It is a subjective measure of perceived usability
    - it should not be your only method
  - It gives no clues about how to improve the score
    - it is not diagnostic
  - It is not possible to make systematic comparisons between two systems and their functionality using SUS

Post-Test Questionnaire: NASA-TLX

- *NASA Task Load indeX (NASA-TLX)*
  - emerged in the 1980s
  - the result of NASA efforts to develop an instrument for measuring the **perceived workload** required by the complex, highly technical tasks of aerospace crew members
- Useful for studying complex products and tasks in high-consequence environments
  - e.g., healthcare, aerospace, military, etc.



NASA-TLX: Questions

- 6 questions on an unlabeled 21-point scale
  - ranging from *Very Low* to *Very High*
- Each question addresses one dimension of the perceived workload:
  - mental demand
  - physical demand
  - time pressure
  - perceived success with the task
  - overall effort level
  - frustration level
- Respondents weigh each one of the questions pertaining to the six categories, to indicate which mattered most to what they were doing

NASA-TLX: Score

- A **complex** instrument to score
- NASA shares a paper and pencil version
  - with instructions
  - https://humansystems.arc.nasa.gov/groups/tlx/tlxpaperpencil.php
- and a free iOS app to compute the score
  - https://itunes.apple.com/us/app/nasa-tlx/id1168110608

Methodology: Think-Aloud

- **Definition:** In a thinking aloud test, you ask test participants to use the system while continuously thinking out loud — that is, simply verbalizing their thoughts as they move through the user interface.
- According to the Nielsen Norman Group, to run a basic thinking aloud usability study, you need to do only 3 things:
  - Recruit representative users
  - Give them representative tasks to perform
  - Shut up and let the users do the talking
- Example: https://www.nngroup.com/articles/thinking-aloud-demo-video/

Methodology: Think-Aloud

- Advantages
  - cheap, you don’t need special equipment
  - simple, it requires little expertise
  - robust, can provide useful insight even with poor facilitators
  - can show how the system is actually used
- Disadvantages
  - unnatural situation
  - subject to biases
  - subject to filtered statements
  - the act of describing may alter task performance (e.g., time-on-task metric)

Methodology: Cooperative Evaluation

- Variation of the think-aloud
- The participant and the facilitator collaborate during the evaluation
  - both can ask each other questions throughout

- Additional advantages
  - less constrained and easier to use
  - user is encouraged to criticize system
  - clarification possible

Equipment

- Any of these can work for an effective usability testing:
  - Laboratory with two or three connected rooms outfitted with audio-visual equipment
  - Room with portable recording equipment
  - Room with no recording equipment, as long as someone is observing the participant and taking notes
  - Remotely, with the participant in a different location (either moderated or unmoderated)

Equipment: Some Material

- Paper and pencil
  - cheap, limited to writing speed
- Audio
  - good for think-aloud
- Video
  - accurate and realistic
  - needs special equipment
  - may be obtrusive

- Computer logging
  - automatic and unobtrusive
  - large amounts of data may be difficult to analyze
- Eye-tracking
  - to track and record eye movements

Mixed use in practice
- audio/video transcription difficult and requires skill
- some automatic support tools available

Sample Scripts

- Sample Usability Testing scripts, with no task described in them, mainly:
  - https://www.sensible.com/downloads/test-script.pdf
  - http://www.lse.ac.uk/intranet/staff/webSupport/guides/archivedWebeditorsHandbook/pdf/script.pdf

Run and Analyze

Usability Testing

Usability Testing: Run

- Get informed consent
  - better in written format
- One person acts as the facilitator and rest of team are observers
  - at least one of the observers must take notes
- Tell each participant:
  - “we are testing our app, not you! Any mistakes are app’s fault, not yours.”
  - **IMPORTANT!**

Usability Testing: Run

- The facilitator should always follow the script, remain neutral, not help the participants, and provide clear instructions
  - tasks can be given in a written form, one at time, … or vocally
- The facilitator must encourage participants to adopt (and explain) the chosen methodologies, at the right moment
  - e.g., how the think-aloud work and for which tasks to use it
- Note-takers take notes of the participant's behavior, comments, errors and completion (success or failure) of each task
- The system is ready to measure **all** the defined criteria

Usability Testing: Analyze

- Analyze collected data to find UI failures and ways to improve
  - e.g., written notes, audio, video, usage logs, …
- Do not forget to consider the collected metrics
  - per task and overall
- Quantitative data can be summarized in, e.g., success rates, task time, error rates, satisfaction questionnaire ratings
- Look for trends and keep a count of problems that occurred across participants
  - e.g., observations about pathways participants took, comments/recommendations, answers to open-ended questions

References

- Alan Dix, Janet Finlay, Gregory Abowd, Russell Beale, Human Computer Interaction, 3rd Edition
  - Chapter 9: Evaluation Techniques
- Ben Shneiderman, Catherine Plaisant, Maxine S. Cohen, Steven M. Jacobs, and Niklas Elmqvist, Designing the User Interface: Strategies for Effective Human-Computer Interaction
  - Chapter 5: Evaluating Interface Design

References

- Beyond the NPS: Measuring Perceived Usability with the SUS, NASA-TLX, and the Single Ease Question After Tasks and Usability Tests
  - https://www.nngroup.com/articles/measuring-perceived-usability/
- John Brooke, SUS - A quick and dirty usability scale, 1986
  - https://hell.meiert.org/core/pdf/sus.pdf
- The Pros and Cons of the System Usability Scale (SUS)
  - https://research-collective.com/blog/sus/

License

- These slides are distributed under a Creative Commons license "**Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)**"

- **You are free to:**
  - **Share** — copy and redistribute the material in any medium or format
  - **Adapt** — remix, transform, and build upon the material
  - The licensor cannot revoke these freedoms as long as you follow the license terms.

- **Under the following terms:**
  - **Attribution** — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
  - **NonCommercial** — You may not use the material for commercial purposes.
  - **ShareAlike** — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.
  - **No additional restrictions** — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

- https://creativecommons.org/licenses/by-nc-sa/4.0/