

Introduction to Human-Computer Interaction

Section 8

Evaluation

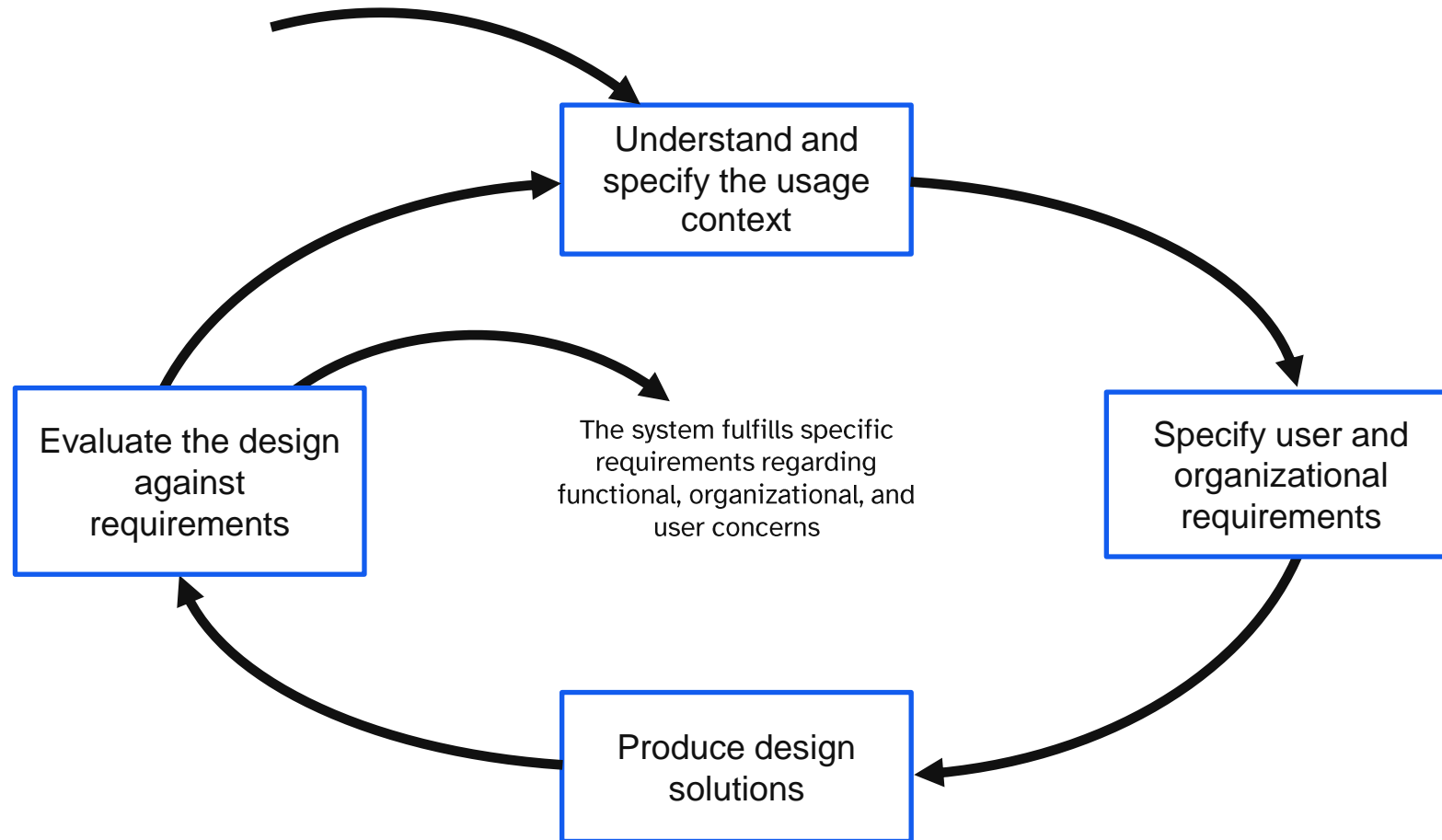


This slide deck was created by Julian Fietkau. It is part of an OER slide collection on the topic “Introduction to Human-Computer Interaction” and belongs to the 2025 version. It is released under the “[Creative Commons Attribution Share-Alike 4.0](#)” license. See section 0, “Information for Educators”, for more details.

https://fietkau.science/teaching/intro_hci

5/30/2025

Reminder: The User-Centered Design Process



Progress from the Previous Sections

Established in section 6:

- How to analyze context and requirements
- The goal of this kind of analysis
- What its results can look like
- How it can be documented

And in section 7:

- How to approach interaction design
- Structured team-based design processes
- Categories of prototypes
- Appropriate types of artifacts for different product maturity levels

➤ How do we conduct evaluations of our artifacts and close the iterative UCD cycle?

Plan for this Section

Continuation of user-centered design: **Evaluations**

- Why evaluate at all?
- Properties of evaluations
- Evaluation methods in detail
 - Questionnaires
 - Interviews
 - Observational studies
 - etc.
- Connecting several methods
 - Usability tests
 - Grounded Theory
- How to plan evaluations

Goal: Good Design

We want the design of our interactive systems to be **successful**. What do we mean by that?

- Reaching functional goals (feature list)?
- Fulfilling project commitments (top down)?
- Observably pleasant user experiences?
- Acceptance by the target audience?

How do we check whether the system we are designing (or have designed) is **usable**?

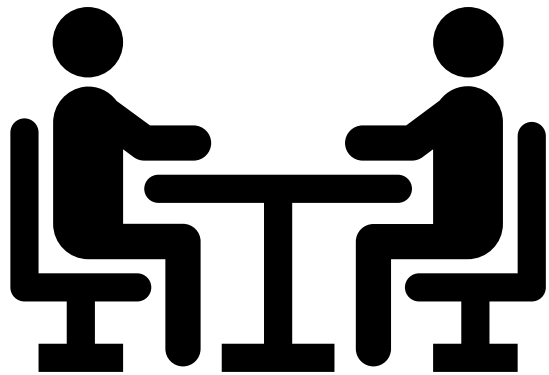
➤ Usability engineering perspective

How do we check whether our ideas and impressions can be **validated / generalized**?

➤ Perspective of HCI as science

Properties of Evaluations

- **Formative or summative:**
 - Formative: early in the design process, intended to find out how the system should be designed
 - Summative: at the end of the design process, intended to find out whether the design process was successful
- **Qualitative or quantitative:**
 - Qualitative: surveyed data is not numerical or scalar in nature
 - Quantitative: surveyed data can be approached numerically / statistically
- **Subjective or objective:**
 - Subjective: based on personal impressions and opinions
 - Objective: based on direct real-world measurements
- **Empirical or analytical:**
 - Empirical: doing experiments and gathering data
 - Analytical: systematic examination of the artifact based on norms and best practices



Evaluation Methods

Questionnaires

- Popular method for collecting subjective data
- Advantages: no inadvertent variance in the question (as you get with interviews, for example), greater comparability
- Disadvantages: Obscures individual details due to rigid structure; if question is unclear, answers are based on personal interpretation (follow-up questions not generally possible)
- Open or closed questions
- Quantitative or qualitative data
- Suitable for pre-test and post-test

Questionnaires: Multiple Choice Questions

- Quantitative data → high comparability, statistic evaluation possible
- Require thoughtful specification of possible answers
- Disadvantage: losing out on personal details and individual backgrounds

I am learning a lot about evaluation methods in this course.

☐ fully agree ☐ partially agree ☐ neither agree nor disagree ☐ partially disagree ☐ fully disagree

Do you understand this question?

☐ yes
☐ no
☐ I don't know

How many siblings do you have?

☐ none ☐ 1 ☐ 2–3 ☐ 4 or more

How many siblings do you have?

Number: ____

What goes on a good pizza?

Multiple answers possible.

☐ Cheese
☐ Pineapple
☐ Mushrooms
☐ Fish sticks

Questionnaires: Likert Scales

Please rate the depicted fish with respect to the following aesthetic criteria.

shapely	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	misshapen
beautiful	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	ugly
attractive	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	revolting



- Special case of multiple choice: placement of a subjective evaluation on a discrete linear scale between two extremes
- Allows extended statistical evaluation (mean, variance)
- Design questions:
 - How many options do you offer?
 - Even or odd number (“neutral” answers allowed)?

Questionnaires: Free-Form Questions

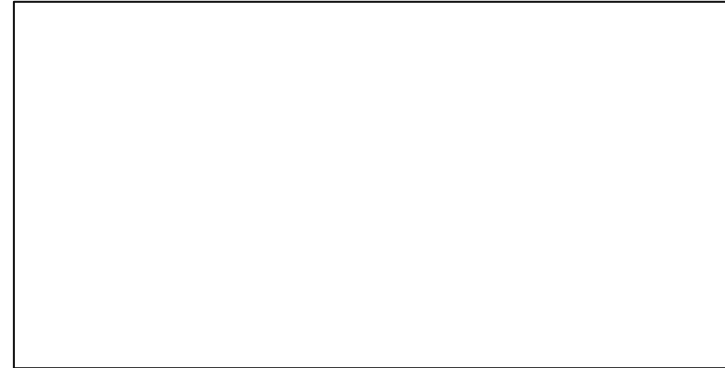
What aspects did you enjoy the most?

- 1.
- 2.
- 3.

What did you enjoy least?

- 1.
- 2.
- 3.

What do we need to improve?



- Qualitative data → more details, more freedom, less comparability, statistical evaluation more difficult
- Open questions are often neglected by test subjects, since they are more time-consuming
- Caution: individual aspects (content of answers, choice of words, handwriting) can undermine anonymity

Questionnaires: Hybrid Questions

Which music genres do you regularly listen to?

Multiple answers possible.

- ☐ Pop
- ☐ Rock
- ☐ Hip Hop
- ☐ Classical
- ☐ Jazz
- ☐ Other: _____

- Can be evaluated like one multiple choice and one free-form question that are related in terms of content
- Free-form “Other:” fields like above are often skipped, may actually fare better as two separate questions

Standardized Questionnaires

- **Validation** of questionnaires: confirmation that items actually measure what they are supposed to measure
 - Validation is labor-intensive and requires several series of tests
- **Standardized questionnaires** are usually already validated multiple times and can be used as measuring instruments for specific variables
- **AttrakDiff**
- **System Usability Scale (SUS)**
- **User Experience Questionnaire (UEQ)**
- ...

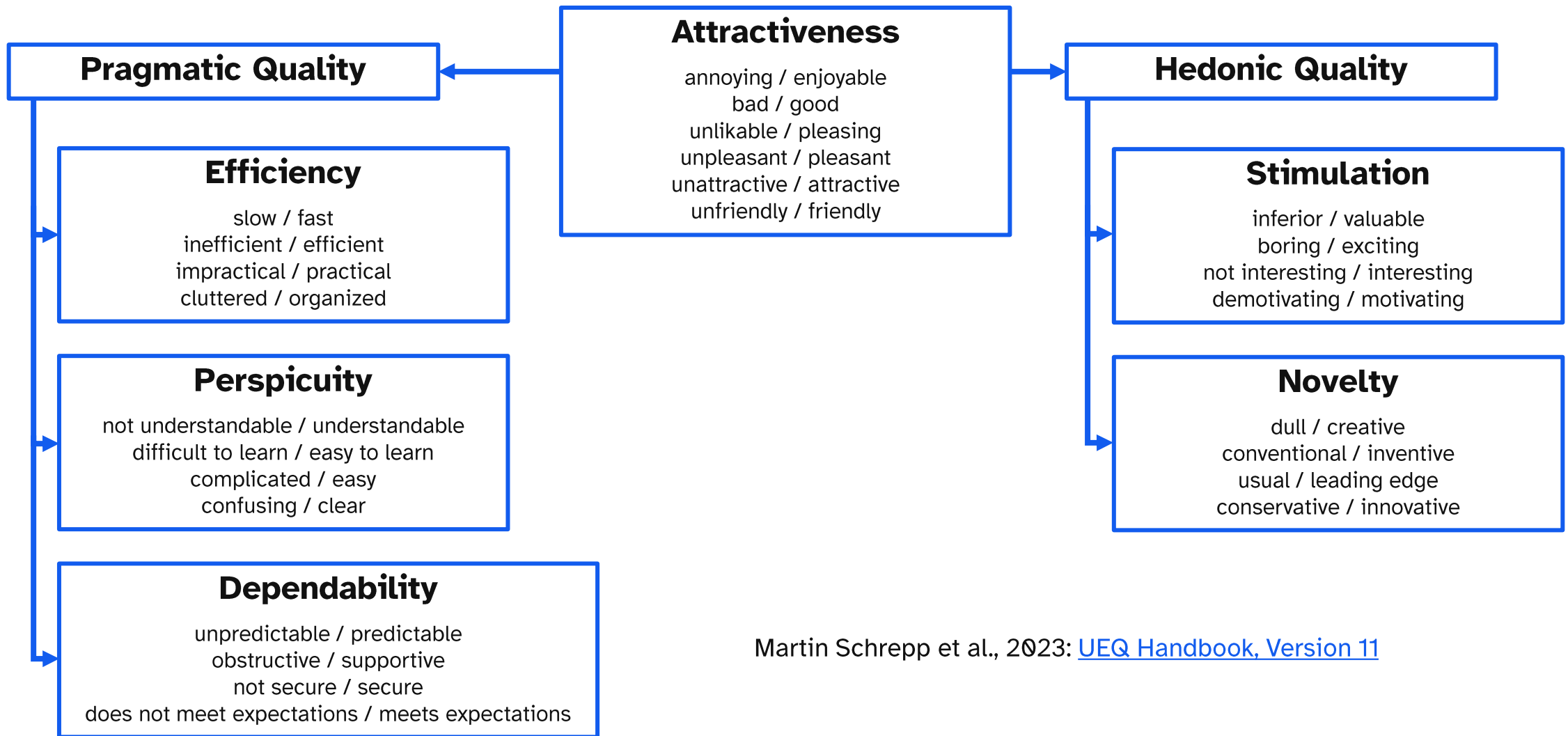
Exercise: User Experience Questionnaire



Please rate _____ by filling out a copy of the **User Experience Questionnaire**.

If the time permits, we will look at the evaluation instruments together later.

User Experience Questionnaire: Measures



Martin Schrepp et al., 2023: [UEQ Handbook, Version 11](#)

Questionnaires: Other Concerns

- Paper or digital?
- On location or online?
- Do you need to detect randomly filled out questionnaires?
 - Which factors raise or lower the danger of this happening?
 - How would you detect problematic cases?
- Further literature on questionnaire design:
 - Bradburn, Sudman & Wansink, 2004: Asking Questions: The Definitive Guide to Questionnaire Design – For Market Research, Political Polls, and Social and Health Questionnaires. 2nd edition. Wiley.
 - Bühner, 2004: Einführung in die Test- und Fragebogenkonstruktion. Pearson Studium.


German

Interviews

Dialogue between experimenter and participant

- Verbal exchange in real time
- Face to face or through some remote communication channel

Types of interviews:

- 
- Structured: questions planned extensively in advance, asked and answered in the same order for all participants
 - Semi-structured: questions planned roughly, spontaneous follow-ups permitted, encouragement of individual focus in the dialogue
 - Unstructured: interview topic planned in advance, questions arise spontaneously based on the exchange

Interviews: Keeping Notes

1. Interviewer keeps their own notes
 - Can delay progression or disturb the flow, is more of an emergency solution if no one else is available to assist
2. Interviewer and minute taker are separate roles
 - Third person is present and takes notes: better solution than 1, notes usually higher quality, potential danger of misguided focus if minute taker is not familiar enough with the topic
3. Audio or video recording
 - Enables a fully detailed post-hoc transcript including the exact wording and tone (and for video, also including gestures and facial expressions), very labor-intensive to evaluate
 - Potentially made easier with AI tools – but: privacy caveats (local processing) and potential for errors rooted in LLM hallucinations
 - Careful with the data: privacy protection (GDPR) vs. data retention obligations in research

Observational Studies

- Participants interact with the system or prototype, experimenters observe
- Fixed task vs. “just try it out”
- Determine in advance what will be recorded
 - Participant’s actions
 - Duration of individual subtasks
 - Mood, emotional state
- Can take place in the lab or in the real environment
 - Or virtually with screen sharing
- Field studies often paired with spontaneous recruitment of test subjects
 - Special case: field studies in public spaces where participants do not know that they are being observed
- As a rule: **resist the temptation to help with problems!**

Observational Studies: Think Aloud

- **Think Aloud** method: participants are asked to express their thoughts **out loud** as comprehensively as they can during the interaction.
 - What they expect or hope to accomplish when they do something
 - What they are looking for
 - How they react mentally when something does not work as expected
 - ...
- Can reveal problems with underlying metaphors and mental processes
- Is a skill that needs to be learned, not everyone is comfortable with it
- As an experimenter: listen, do not judge (do not indicate agreement or disagreement), and certainly do not contradict
- Variation: two participants tell each other what they think

Exercise: Think Aloud Test



Carry out a short think-aloud experiment as a team of two.

You can use your paper prototype from the last section or, alternatively, use an interactive software (e.g. a website that you know well enough) that the other person is not familiar with.

Focus Groups

- Method originally designed for market research
- A focus group consists of six to twelve people who all have a certain common demographic characteristic
- Focus groups conceivable for: students, pensioners, single mothers, dog owners, heavy metal listeners, ...
- Guided discussion on a specific topic
 - Unlike a 1:1 interview, **group dynamics** emerge
- Objective: collection of representative personal attitudes of the demographic group
- Differentiation criterion from other kinds of group interviews: selection and contrasting of demographic characteristics

Usage Data Logging

- Interactive systems can log usage data internally:
 - What was clicked how often?
 - Which interactions led to abandonment, which led to further successful interactions?
- Which external factors influence usage?
 - Day of the week? Time of day? Location? Weather? ...
- Evaluation also possible over longer periods of time
- No resource cost from continuous human observation

Eye Tracking and Other Sensors

- Recording **physical aspects** of the subject's behavior: eye or hand movement, movement of the person in space, ...
- Sensors used to be clunky, heavy, limited to brief use in the laboratory; today they are often barely noticeable
- **Objective data:** enabling the investigation of differences between stated and actual behavior
 - Deviations of the test subjects from their own ideals
 - Unconscious aspects of attention

Eye Tracking: Evaluation Tool (Example)

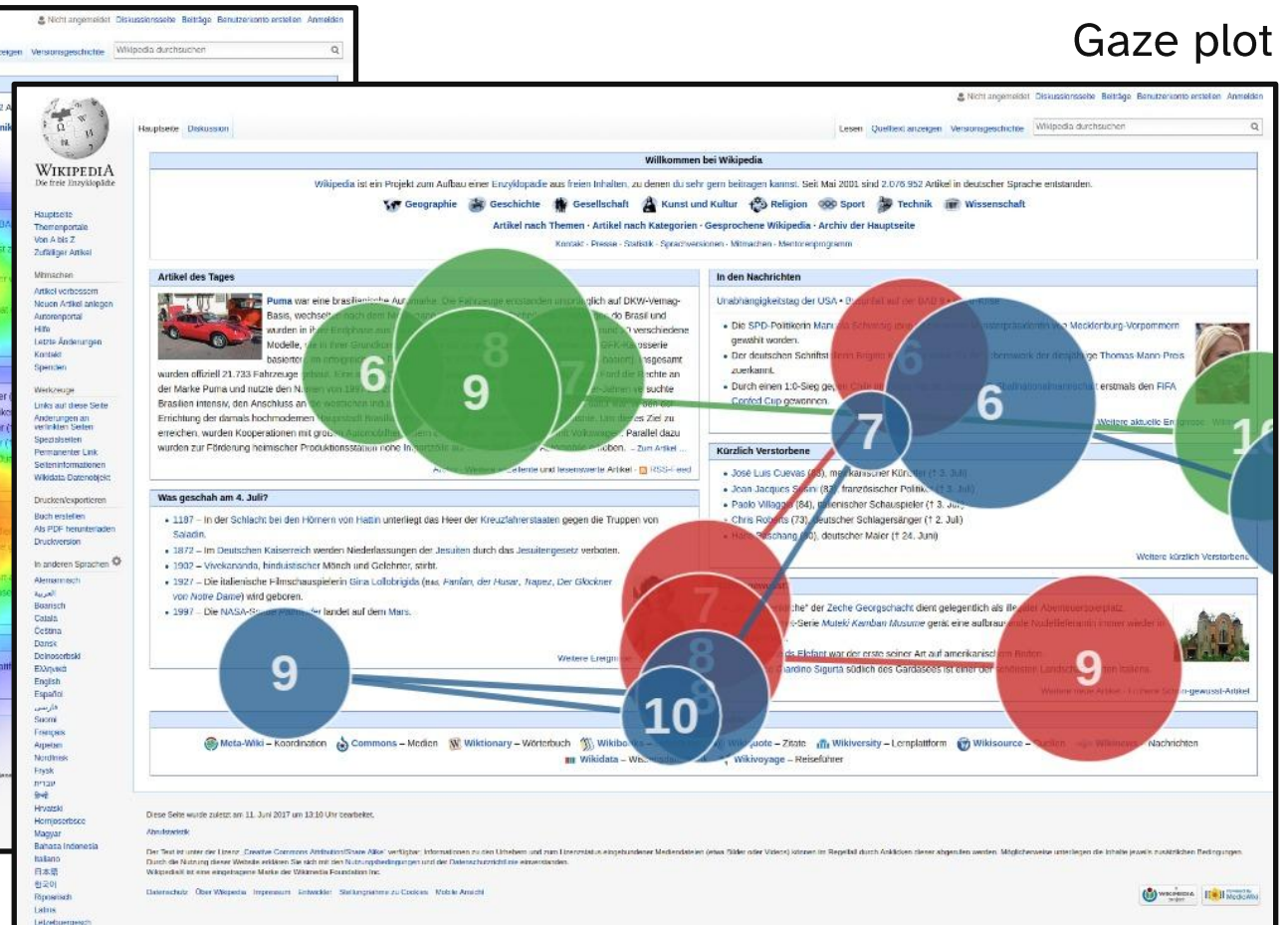


Chronos Vision GmbH, 2009: [Eye Tracking Device GUI](#), Wikimedia Commons / [CC-BY-SA 3.0](#)

Eye Tracking: Visualizations



Heat map



Gaze plot

User: Tschneidr, 2017: [Eyetracking heat map Wikipedia](#), [Gaze plot eye tracking on Wikipedia with 3 participants](#), Wikimedia Commons / [CC-BY-SA 4.0](#)

Usability Tests

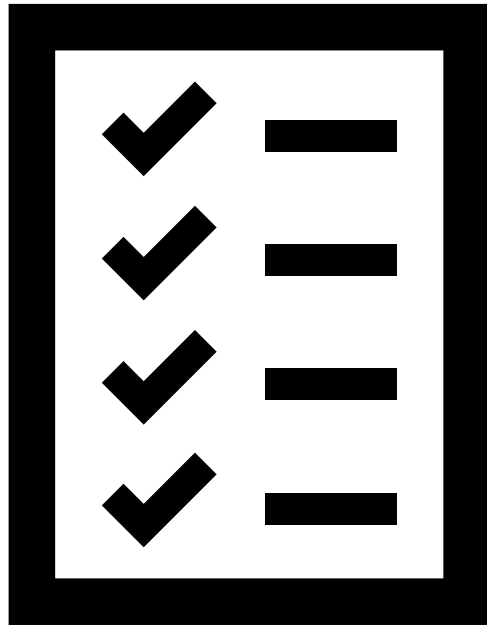
- In practice, usability tests often consist of several individual empirical methods
 - e.g. questionnaire, then observation, then interview
- Recommended: test run of the complete plan with colleagues before inviting real participants
 - Check whether equipment, artifacts, procedures work
 - Less experienced experimenters: practicing the procedures, uncovering gaps in the planning
 - Gather experience regarding the total duration of a run
- Do all participants use the same version of the system (maximum comparability) or are problems fixed between the individual runs (maximum progress in results)?

Grounded Theory

- Methodology for gaining knowledge from a large amount of empirical (generally qualitative) data
- Step-by-step development of a theoretical understanding from the data, attempt to avoid influence by the experimenters' preconceptions
- Coding of the available data: working out patterns and categories, linking through logical connections
- Theory is determined “bottom up” from the data instead of being prescribed “top down”, e.g. as in simple hypothesis testing
- Suitable data sources: interviews, observations, text mining, ...

Heuristic Evaluation / Expert Analysis

- Evaluation of an interactive system based on a specific set of rules or heuristics
- Performed by one or more usability experts who were not involved in the design process
- Possible evaluation criteria:
 - ISO 9241-110
 - 8 golden rules according to Shneiderman
 - Nielsen's heuristics
 - ...
 - see section 4
- Provides different answers than an empirical evaluation and is not a drop-in replacement for one



Planning Evaluations

Setting an Evaluation Goal

Validation of an idea or concept

Comparison of two products or versions

Assessing development progress

Checking requirements fulfillment

...

Evaluation Strategies

High level (**Strategy**)

- What is to be achieved by the evaluation?
- What external restrictions are in place?
- What is the target user group, who is eligible as a participant?
- What resources are available?
- ...

Low level (**Plan**)

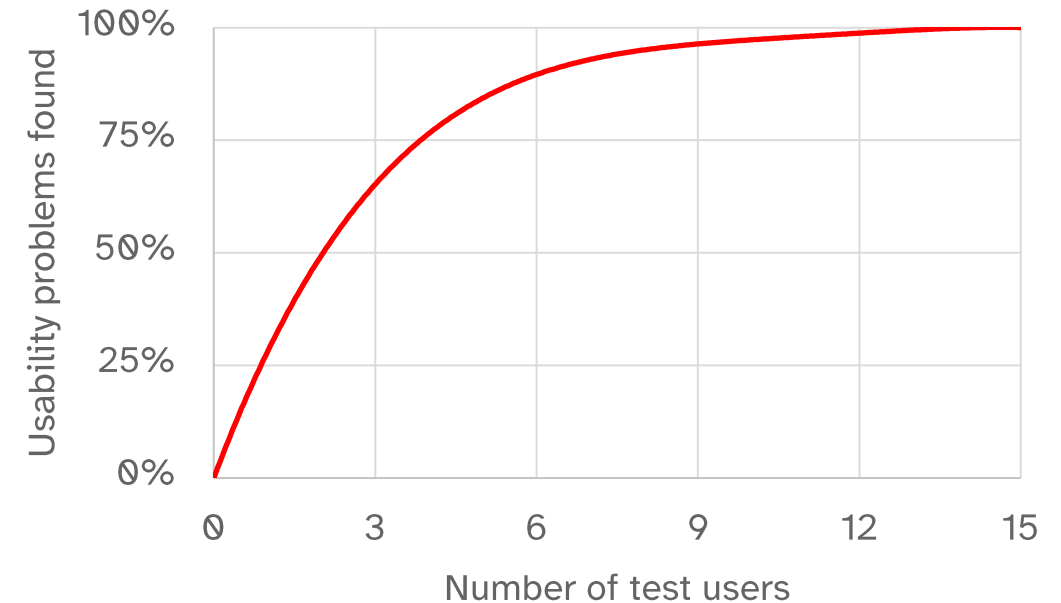
- Which evaluation artifacts need to be prepared?
- Where and when do the sessions take place?
- How are participants recruited, how are they assigned an appointment?
- Buffer time? Catering (water, snacks)? Hygiene?
- ...

Method Choice: Parameters

- Required data
 - Statistically significant demographic statements?
 - Detailed individual opinions?
 - In-depth discursive insights?
- Resources
 - Time
 - Money
 - Hardware
 - Availability of test subjects (target demographic!)
- Stage of the design process
 - Ideation phase
 - Lo-fi prototypes
 - Tests on the real system

Number of Participants

- The infamous **n**
- Depends on the goal and methodology
 - Surfacing usability issues in a product: $n \leq 15$ may well cover most of the important ground
 - Statistical hypothesis testing based on statements regarding e.g. national populations: might need to go with $n > 1000$



Jakob Nielsen, 2000: [Why You Only Need to Test with 5 Users](#)

Summary of the Evaluation Design Process

Evaluation goal

leads to

Method choice

leads to

Schedule

Recommendations for Interacting with Participants

- Treat your participants with respect.
- Set yourself the goal to make the evaluation itself “user-centered” as well, not just the system.
- Inform the participants about the evaluation process in advance.
- The system is being evaluated, not the participant.

For More, See *HCI Lecture “Evaluation & Experiments”*

Introduction to Evaluation and Experiments

An introduction into analytical and empirical evaluations in HCI

Cognitive Walkthrough

Understanding the cognitive walkthrough as evaluation method

Heuristic Evaluation

Principles and heuristics as analytical evaluation method

Usability Testing

Conducting usability tests with real users

Grounded Theory

Qualitative data analysis

Empirical Research

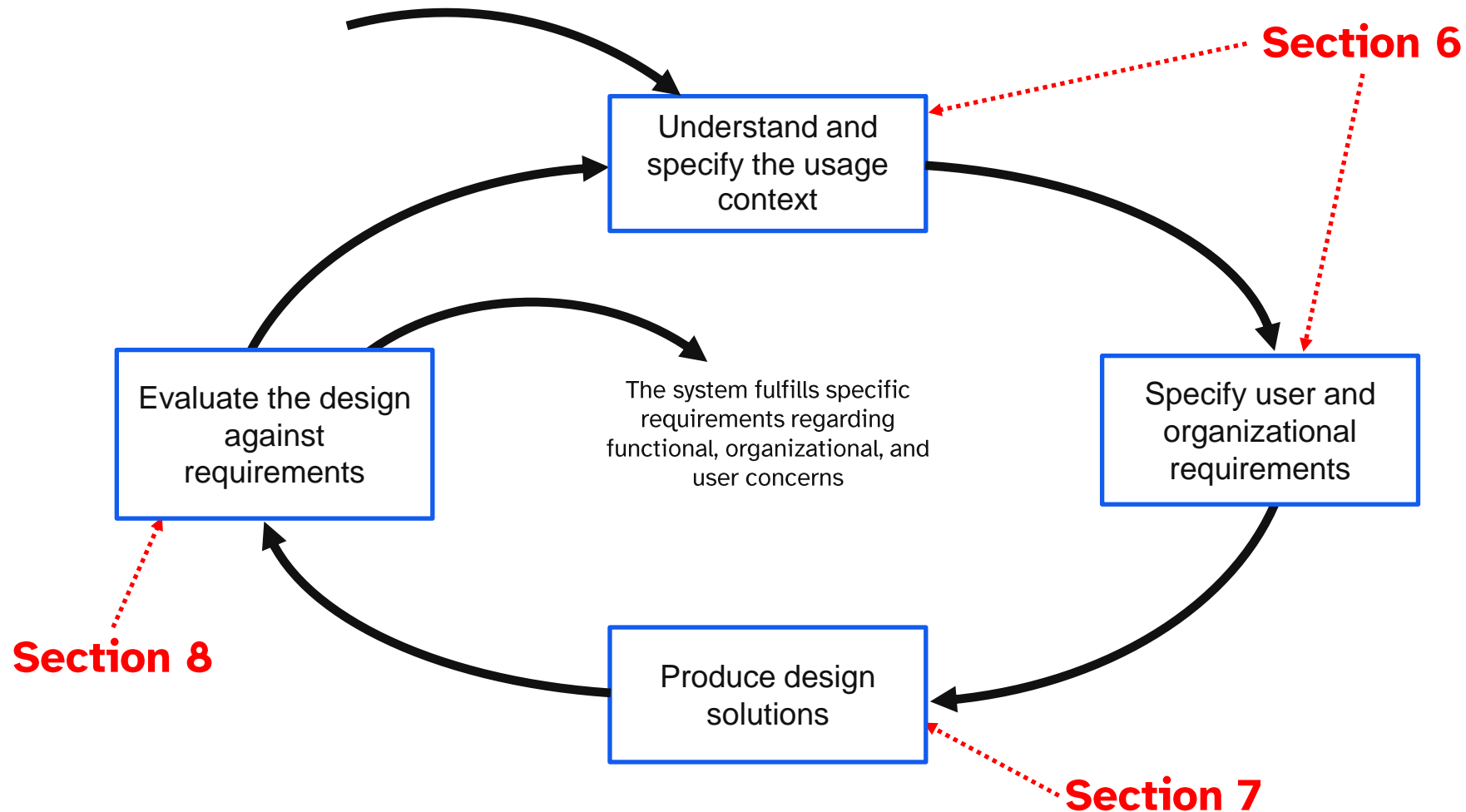
Basics and principles of empirical user studies

Hypothesis Testing

Testing hypothesis for statistical significance

Albrecht Schmidt et al., 2020: [HCI Lecture: Material for Teaching Human-Computer Interaction, Evaluation & Experiments](#)

Reminder: The User-Centered Design Process



Overview

1	Introduction and Overview
2	Basics of Cognition
3	Perception and Communication
4	Guidelines for Interaction Design
5	The Usability Engineering Process
6	Context Analysis
7	Design and Prototyping
8	Evaluation
9	Interaction Paradigms
10	Computer-Supported Cooperative Work
11	Accessibility
12	Information architecture and data visualization
13	Visual Design
14	HCI over Time
15	Professional Values and Ethics



Part II: Methods

What do we do? In what order?
What are our success criteria?



Review: This Section's Topics

Evaluation methods: questionnaires, interviews, observational studies, focus groups, usage data logging, eye tracking, usability tests, grounded theory, heuristic evaluation

Evaluation planning: what can evaluation accomplish, evaluation goals, planning and strategy, method selection

Questions for a quick self-assessment:

1. What are properties of Likert scales?
2. What makes an interview semi-structured?
3. What are typical parts of a usability test?
4. What is the difference between the concepts “quantitative” and “objective”?
5. How many participants does an evaluation need? What does the answer depend on?