

Einführung in Mensch-Computer-Interaktion

Abschnitt 8

Evaluation

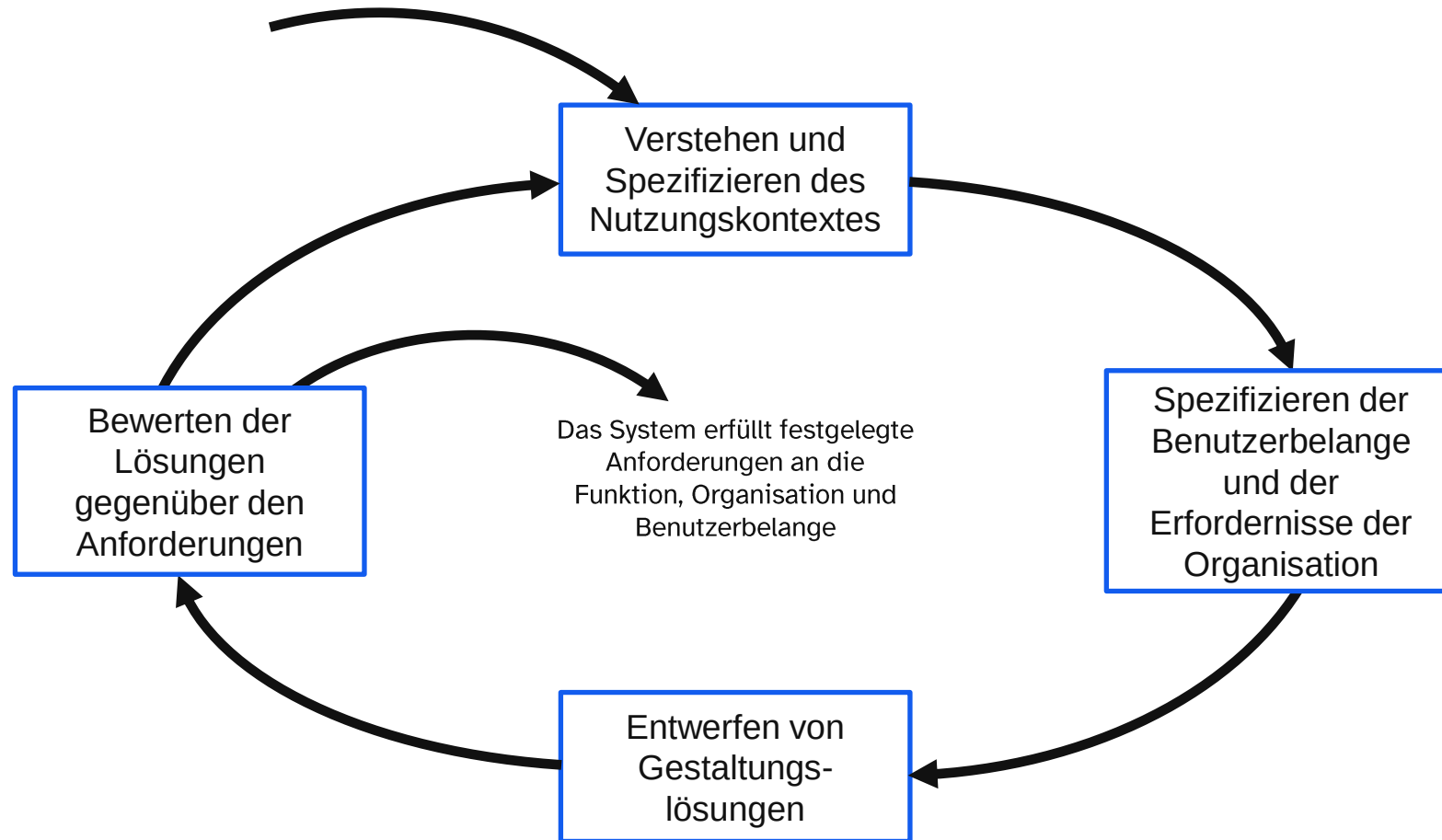


Dieser Foliensatz wurde von Julian Fietkau erstellt und ist Teil einer OER-Foliensammlung zum Thema „Einführung in Mensch-Computer-Interaktion“ in der Version von 2025. Er ist unter der Lizenz „Creative Commons Attribution Share-Alike 4.0“ freigegeben. Eine ausführliche Erläuterung ist in Abschnitt 0 „Information für Lehrende“ zu finden.

https://fietkau.science/teaching/intro_hci

30.05.2025

Zur Erinnerung: Ablauf User-Centered Design



Zwischenstand nach den letzten zwei Abschnitten

Bekannt aus Abschnitt 6:

- Wie Kontext- und Anforderungsanalyse funktioniert
- Was das Ziel dieser Analysen ist
- Wie ihre Ergebnisse aussehen
- Welche Dokumentationsformen es gibt

Und aus Abschnitt 7:

- Wie Interaktionsdesign angegangen werden kann
 - Wie man dabei als Team strukturiert vorgeht
 - Welche Arten von Prototypen es gibt
 - Was für Artefakte man für welchen Reifegrad des Produkts verwenden sollte
- Wie schreiten wir in die Evaluation unserer Artefakte und schließen den iterativen UCD-Kreislauf?

Plan für diesen Abschnitt

Fortsetzung nutzerzentrierte Gestaltung: **Evaluationen**

- Warum überhaupt evaluieren?
- Welche Eigenschaften haben Evaluationsverfahren?
- Evaluationsmethoden im Detail
 - Fragebögen
 - Interviews
 - Beobachtungsstudien
 - etc.
- Verknüpfungen mehrerer Methoden
 - Usability-Tests
 - Grounded Theory
- Planung von Evaluationen

Ziel: Gute Gestaltung

Wir wollen, dass die interaktiven Systeme die wir bauen **gut gelingen**. Was heißt das?

- Erreichen funktionaler Ziele (Feature-Umfang)?
- Erfüllung von Projektanforderungen (top down)?
- Beobachtbare angenehme Nutzungserfahrung?
- Akzeptanz durch die Zielgruppe?

Wie stellen wir fest, ob das System das wir gestaltet haben (oder gerade gestalten) **gebrauchstauglich** ist?

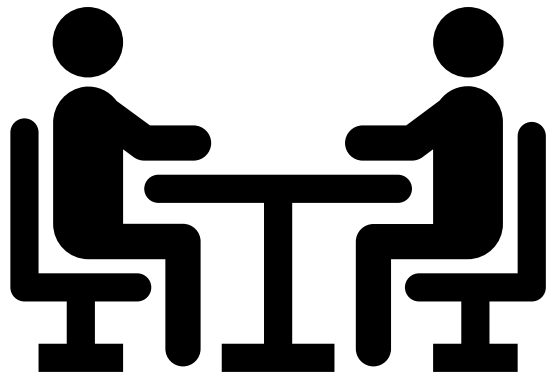
➤ Usability-Engineering-Perspektive

Wie stellen wir fest, ob unsere eigenen Ideen und Eindrücke **überprüfbar / allgemeingültig** sind?

➤ Perspektive der MCI als Wissenschaft

Eigenschaften von Evaluationsverfahren

- **Formativ** oder **summativ**:
 - **Formativ**: früh im Gestaltungsprozess, soll herausfinden wie das System gestaltet werden soll
 - **Summativ**: am Ende des Gestaltungsprozesses, soll herausfinden ob der Gestaltungsprozess erfolgreich war
- **Qualitativ** oder **quantitativ**:
 - **Qualitativ**: erhobene Daten sind nicht-numerischer Natur
 - **Quantitativ**: erhobene Daten sind numerisch bzw. statistisch auswertbar
- **Subjektiv** oder **objektiv**:
 - **Subjektiv**: basierend auf persönlichen Eindrücken und Meinungen
 - **Objektiv**: basierend auf realweltlichen Messungen
- **Empirisch** oder **analytisch**:
 - **Empirisch**: Durchführung von Experimenten, Erhebung von Daten
 - **Analytisch**: Systematische Betrachtung des Untersuchungsgegenstands anhand von Normen und Best Practices



Evaluationsmethoden

Fragebögen

- Beliebte Methode zur Erhebung von subjektiven Daten
- Vorteile: keine versehentliche Varianz in der Fragestellung (wie z.B. bei Interviews), stärkere Vergleichbarkeit
- Nachteile: Verdeckung individueller Details durch rigide Struktur; falls Frage unklar, erfolgt Beantwortung auf Basis der persönlichen Interpretation (Nachfragen nicht immer möglich)
- Offene oder geschlossene Fragen
- Quantitative oder qualitative Daten
- Geeignet für Pre-Test und Post-Test

Fragebögen: Multiple-Choice-Fragen

- Quantitative Daten → hohe Vergleichbarkeit, statistische Auswertbarkeit
- Erfordern gute Vorab-Spezifikation der benötigten Antwortmöglichkeiten
- Nachteil: Verlust möglicher Details und individueller Hintergründe

Ich lerne durch diese Vorlesung Evaluationsmethoden besser kennen.

☐ stimme voll zu ☐ stimme weitgehend zu ☐ stimme halbwegs zu ☐ stimme kaum zu ☐ stimme nicht zu

Verstehen Sie diese Frage?

- ☐ ja
☐ nein
☐ weiß ich nicht

Wie viele Geschwister haben Sie?

☐ keine ☐ 1 ☐ 2–3 ☐ 4 oder mehr

Wie viele Geschwister haben Sie?

Anzahl: ____

Was gehört auf eine gute Pizza?

Mehrfachantwort möglich.

- ☐ Ananas
☐ Champignons
☐ Fischstäbchen
☐ Käse

Fragebögen: Likert-Skalen

Bitte bewerten Sie den abgebildeten Fisch hinsichtlich der folgenden ästhetischen Kriterien.

formschön	☐	☐	☐	☐	☐	☐	unförmig
hübsch	☐	☐	☐	☐	☐	☐	hässlich
anziehend	☐	☐	☐	☐	☐	☐	abstoßend



- Sonderfall von Multiple Choice: Platzierung einer subjektiven Bewertung auf einer diskreten linearen Skala zwischen zwei Extrema
- Erlaubt erweiterte statistische Auswertung (Mittelwert, Varianz)
- Entwurfsfragen:
 - Wie viele Antwortmöglichkeiten sehen Sie vor?
 - Gerade oder ungerade Anzahl? („mittel“ als Antwort erlauben?)

Fragebögen: Offene Fragen

Was hat Ihnen am besten gefallen?

- 1.
- 2.
- 3.

Was am schlechtesten?

- 1.
- 2.
- 3.

Was sollte noch verbessert werden?

- Qualitative Daten → mehr Details, mehr Freiraum, weniger Vergleichbarkeit, statistische Auswertung erschwert
- Offene Fragen durch Versuchspersonen oft vernachlässigt, da zeitaufwändiger
- Vorsicht: Individuelle Aspekte (Inhalt der Antworten, Wortwahl, Handschrift) können die Anonymität aushebeln

Fragebögen: Hybride Fragen

Welche Musikgenres hören Sie regelmäßig?

Mehrfachantwort möglich.

- ☐ Pop
- ☐ Rock
- ☐ Hip Hop
- ☐ Klassik
- ☐ Schlager
- ☐ Sonstige: _____

- Auswertbar wie eine geschlossene und eine offene Frage, die inhaltlich zusammen gehören
- Offene Ergänzungsfelder wie oben normalerweise nur selten ausgefüllt, nacheinander folgende Fragen ggf. besser

Standardisierte Fragebögen

- **Validierung** von Fragebögen: Bestätigung dass Items tatsächlich das messen, was sie messen sollen
 - Validierung ist arbeitsintensiv und erfordert mehrere Versuchsreihen
- **Standardisierte Fragebögen** sind zumeist bereits vielfältig validiert und können als Messinstrumente für bestimmte Größen verwendet werden
- **AttrakDiff**
- **System Usability Scale (SUS)**
- **User Experience Questionnaire (UEQ)**
- ...

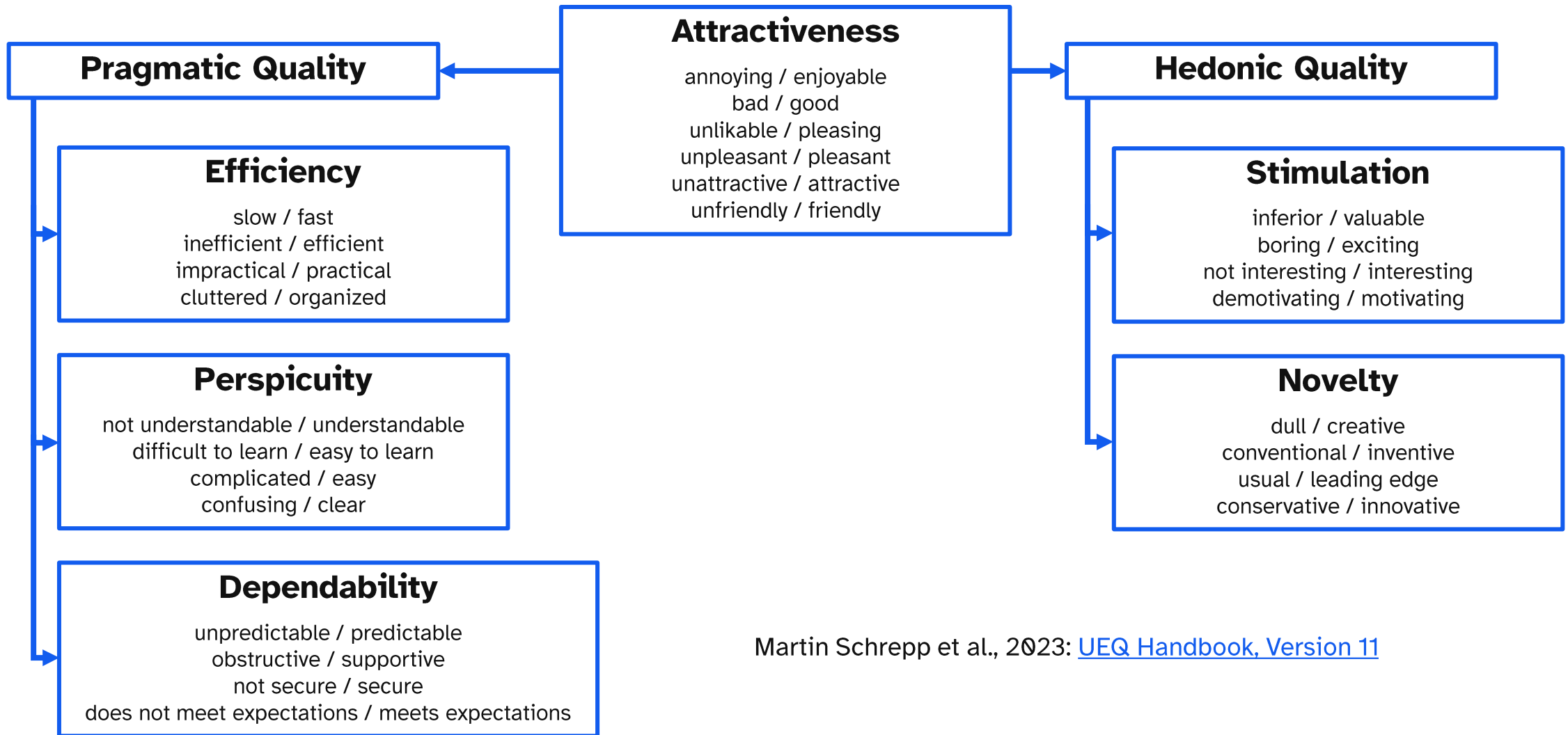
Übung: User Experience Questionnaire



Bewerten Sie _____, indem Sie ein Exemplar des **User Experience Questionnaire** ausfüllen.

Wenn die Zeit es zulässt, schauen wir uns noch gemeinsam die Auswertung an.

User Experience Questionnaire: Messgrößen



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

Fragebögen: Weitere Punkte

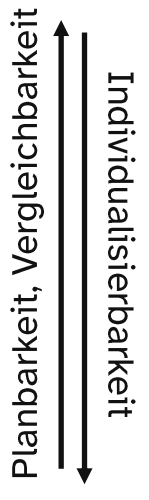
- Papier oder digital?
- Vor Ort oder online?
- Erkennung zufällig angekreuzter Antworten nötig?
 - Was fördert, was hemmt diese Gefahr?
 - Wie erkennt man Problemfälle?
- Weitere Literatur zur Fragebogengestaltung:
 - Bühner, 2004: Einführung in die Test- und Fragebogenkonstruktion. Pearson Studium.
 - Bradburn, Sudman & Wansink, 2004: Asking Questions: The Definitive Guide to Questionnaire Design – For Market Research, Political Polls, and Social and Health Questionnaires. 2. Auflage. Wiley.

Interviews

Gespräch zwischen Experimentator*in und Versuchsperson

- Mündlicher Austausch in Echtzeit
- In Präsenz oder via Telekommunikations-Kanal

Arten von Interviews:

- 
- Strukturiert: Fragenkatalog vorher ausführlich geplant, mit allen Versuchspersonen in der gleichen Reihenfolge abgearbeitet
 - Semistrukturiert: Fragenkatalog grob geplant, spontane Folgefragen erlaubt, Ermunterung zu individuellen Schwerpunkten im Gespräch
 - Unstrukturiert: Interviewthemen vorher festgelegt, Fragen ergeben sich spontan aus dem Gesprächsverlauf

Interviews: Protokollierung

1. Interviewer*in macht selbst Notizen
 - Verzögert den Ablauf, stört die Konzentration, nur Notlösung falls keine zweite Person zur Verfügung steht die helfen kann
2. Interviewer*in und Protokollant*in sind getrennte Rollen
 - Dritte Person sitzt dabei und protokolliert: bessere Lösung als 1, Notizen i.d.R. von höherer Qualität, potenzielle Gefahr einer ungeeigneten Schwerpunktsetzung falls Protokollant*in nicht gut genug ins Thema eingearbeitet
3. Audio- oder Videoaufzeichnung
 - Ermöglicht nachträgliche Transkription inkl. vollständigem Wortlaut (Video: zusätzlich Gestik und Mimik), sehr arbeitsintensiv in der Auswertung
 - Potenzielle Erleichterung durch KI-Tools – aber: Achtung der Privatsphäre (lokale Verarbeitung)
 - Vorsicht mit den Daten: Schutz der Privatsphäre (DSGVO) vs. Datenhaltungspflicht in der Forschung

Beobachtungsstudien

- Versuchspersonen interagieren mit dem System bzw. dem Prototypen, Experimentator*innen beobachten
- Fest vorgegebene Aufgabe vs. „einfach mal ausprobieren“
- Vorher festlegen, was protokolliert wird
 - Aktionen der Versuchsperson
 - Dauern einzelner Teilaufgaben
 - Laune, emotionaler Zustand
- Kann im Labor oder im realen Kontext stattfinden
 - Oder auch virtuell mit Screen-Sharing
- Feldstudien oft gepaart mit spontaner Rekrutierung von Versuchspersonen
 - Sonderfall: Feldstudien im öffentlichen Raum, bei denen Versuchspersonen nicht wissen, dass sie beobachtet werden
- In aller Regel: **der Versuchung widerstehen, bei Problemen zu helfen!**

Beobachtungsstudien: Think Aloud

- „**Think Aloud**“-Methode: Versuchspersonen werden gebeten, ihre Gedanken während der Interaktion möglichst umfassend **auszusprechen**.
 - Was sie erwarten bzw. sich erhoffen, wenn sie etwas tun
 - Wonach sie suchen
 - Wie sie gedanklich reagieren, wenn etwas nicht funktioniert wie erwartet
 - ...
- Kann Probleme mit unterliegenden Metaphern und mentalen Prozessen zu Tage fördern
- Ist eine Fähigkeit, die erlernt werden muss – nicht jede Versuchsperson fühlt sich damit wohl
- Als Experimentator*in: zuhören, nicht werten (Zustimmung oder Ablehnung nicht anmerken lassen), schon gar nicht widersprechen
- Variante: zwei Versuchspersonen erzählen sich gegenseitig, was sie denken

Übung: Think-Aloud Test



Führen Sie mit zwei Personen eine beispielhafte Think-Aloud-Aufgabe durch.

Sie können dafür Ihren Papierprototypen vom letzten Abschnitt verwenden, oder alternativ eine interaktive Software (bspw. eine Ihnen gut genug bekannte Webseite) testen lassen, die die andere Person nicht gut kennt.

Fokusgruppen

- Methode ursprünglich aus der Marktforschung
- Fokusgruppe besteht aus sechs bis zwölf Personen, welche alle eine bestimmte gemeinsame demografische Eigenschaft haben
- Fokusgruppen denkbar für: Studierende, Rentner*innen, alleinerziehende Mütter, Hundehalter*innen, Heavy-Metal-Hörende, ...
- Angeleitete Diskussion zu einem bestimmten Thema
 - Anders als beim 1:1-Interview entsteht **Gruppendynamik**
- Ziel: Sammlung repräsentativer persönlicher Einstellungen der demografischen Gruppe
- Abgrenzungskriterium zum allgemeinen Gruppeninterview: Wahl und Kontrastierung demografischer Merkmale

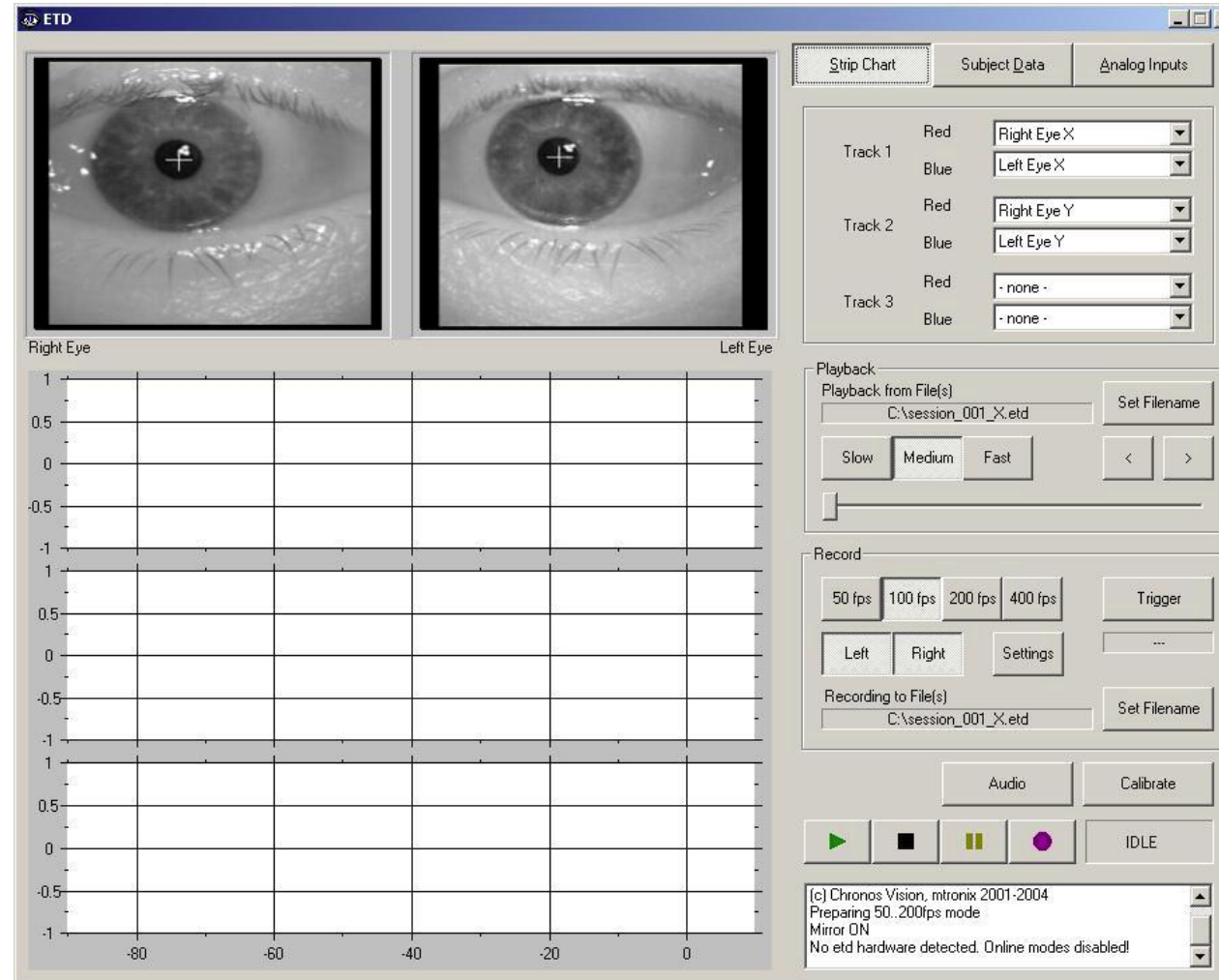
Nutzungsdaten-Logging

- Interaktive Systeme können Nutzungsdaten intern protokollieren:
 - Was wurde wie oft angeklickt?
 - Welche Interaktionen führten zu Abbrüchen, welche zu weiteren erfolgreichen Interaktionen?
- Welche externen Faktoren beeinflussen die Nutzung?
 - Wochentag? Uhrzeit? Standort? Wetter? ...
- Auswertung auch über längere Zeiträume möglich
- Kein Aufwand durch menschliche Beobachtung

Eye Tracking und andere Sensorik

- Aufzeichnung **physischer Aspekte** des Verhaltens der Versuchspersonen: Bewegungen von Augen oder Händen, Bewegung der Person im Raum, ...
- Sensoren früher noch klobig, schwer, auf kurze Einsätze im Labor beschränkt; heute oft kaum noch sichtbar
- **Objektive Daten:** Ermöglichung der Untersuchung von Unterschieden zwischen angegebenem und realem Verhalten
 - Abweichungen der Versuchspersonen von ihren eigenen Idealvorstellungen
 - Unbewusste Aspekte der Aufmerksamkeit

Eye Tracking: Werkzeug zur Auswertung (Beispiel)

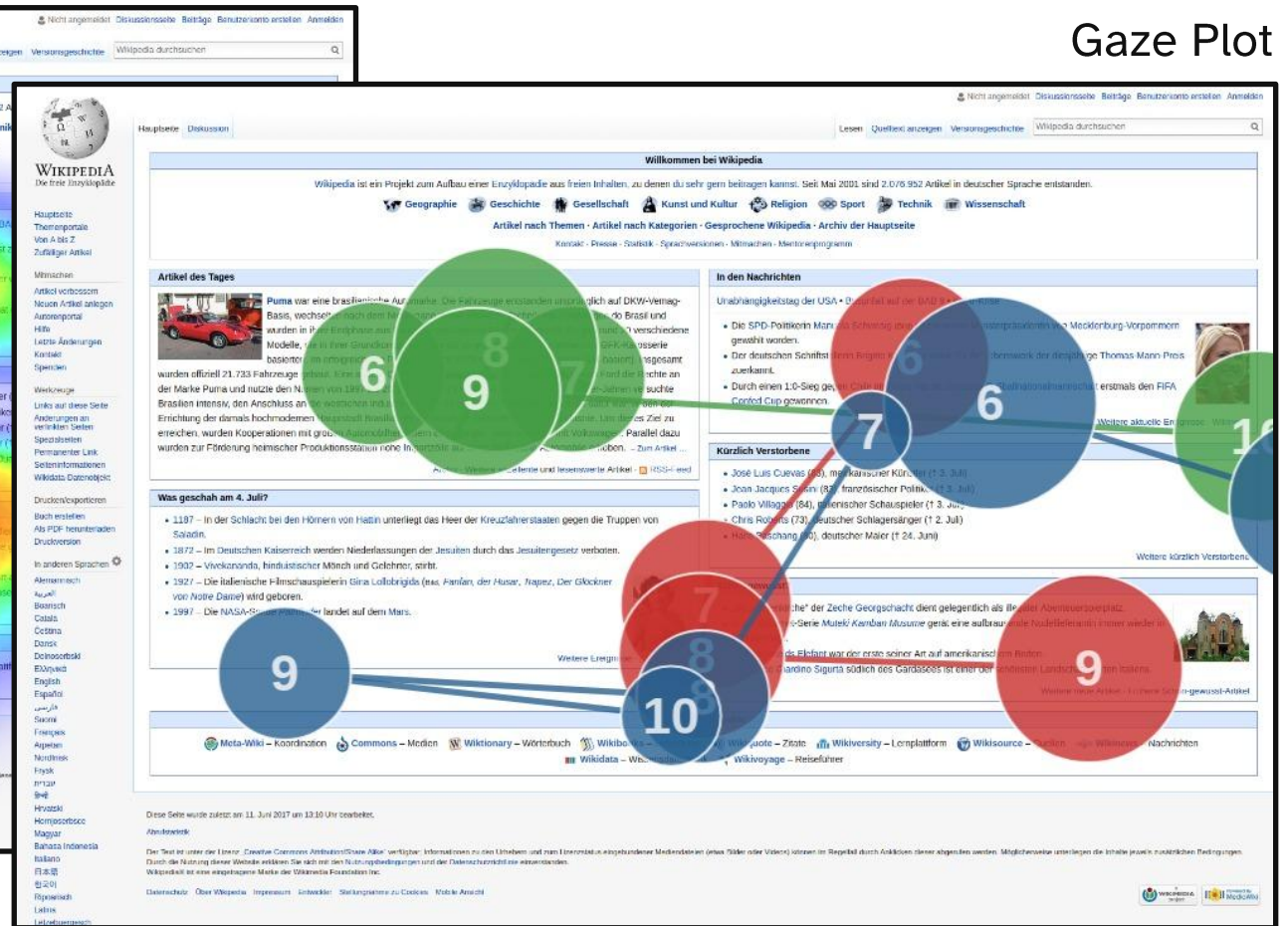


Chronos Vision GmbH, 2009: [Eye Tracking Device GUI](#), Wikimedia Commons / [CC-BY-SA 3.0](#)

Eye Tracking: Visualisierungen



Heatmap



Gaze Plot

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

Usability-Tests

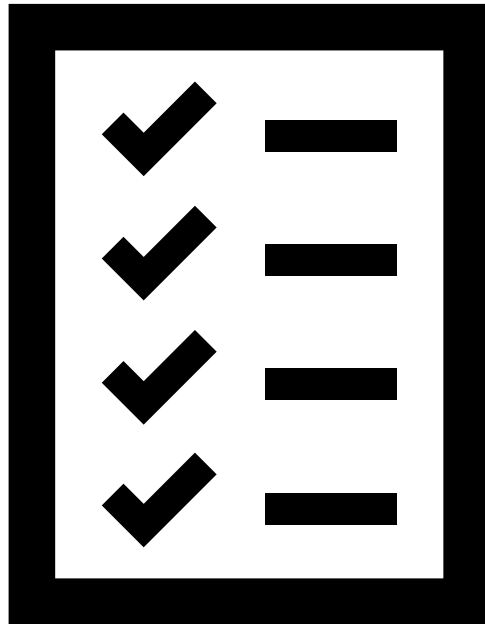
- Usability-Tests setzen sich in der Praxis häufig aus mehreren einzelnen empirischen Methoden zusammen
 - z.B. Fragebogen, dann Beobachtung, dann Interview
- Empfehlenswert: Testdurchlauf des vollständigen Plans mit Kollegen, bevor reale Versuchspersonen eingeladen werden
 - Prüfung ob Equipment, Artefakte, Abläufe funktionieren
 - Weniger erfahrene Experimentatoren: Üben der Abläufe, Aufdecken von Lücken in der Planung
 - Erfahrungswerte sammeln bzgl. der Gesamtdauer eines Durchlaufs
- Testen alle Versuchspersonen die gleiche Version des Systems (maximale Vergleichbarkeit) oder werden Probleme zwischen den einzelnen Durchläufen behoben (maximaler Ergebnisfortschritt)?

Grounded Theory

- Methodik für die Erkenntnisgewinnung aus einer großen Menge empirischer (meist qualitativer) Daten
- Schrittweise Entwicklung eines Theorieverständnisses aus der Datenbasis heraus, Vermeidung von Einfluss durch Vorurteile der Experimentatoren
- Kodierung der vorliegenden Daten: Herausarbeitung von Mustern und Kategorien, Verknüpfung durch logische Zusammenhänge
- Theorie wird „bottom up“ aus den Daten ermittelt statt „top down“ vorgegeben zu werden wie z.B. beim einfachen Hypothesentest
- Geeignete Datenquellen: Interviews, Beobachtungen, Text Mining, ...

Heuristische Evaluation / Evaluation durch Experten

- Bewertung eines interaktiven Systems anhand eines bestimmten Satzes an Regeln oder Heuristiken
- Durchgeführt durch einen oder mehrere Usability-Experten, welche nicht an der Gestaltung beteiligt waren
- Mögliche Bewertungskriterien:
 - ISO 9241-110
 - 8 goldene Regeln nach Shneiderman
 - Nielsens Heuristiken
 - ...
 - siehe Abschnitt 4
- Liefert andere Antworten als eine empirische Evaluation und kann diese nicht gleichwertig ersetzen



Planung von Evaluationen

Evaluationsziel festlegen

Validierung einer Idee / eines Konzepts

Vergleich von zwei Produkten / Versionen

Entwicklungsfortschritt prüfen

Erfüllung der Anforderungen bewerten

...

Evaluationsstrategie

High Level (**Strategie**)

- Was soll durch die Evaluation erreicht werden?
- Welche externen Einschränkungen gibt es?
- Was ist die Zielgruppe, wer kommt als Versuchsperson in Frage?
- Welche Ressourcen stehen zur Verfügung?
- ...

Low Level (**Plan**)

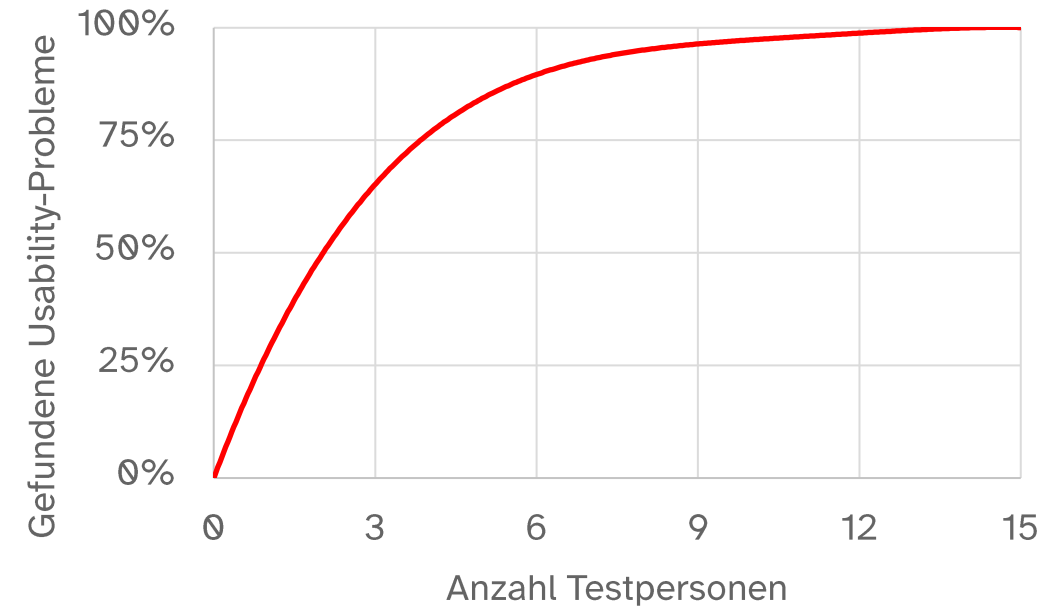
- Welche Evaluationsartefakte müssen vorbereitet werden?
- Wo und wann finden die Sitzungen statt?
- Wie werden Versuchspersonen gewonnen, wie bekommen sie einen Termin zugeordnet?
- Pufferzeit? Verpflegung (Wasser, Snacks)? Hygiene?
- ...

Methodenwahl: Parameter

- Benötigte Daten
 - Statistisch belastbare demografische Aussagen?
 - Ausführliche Einzelmeinungen?
 - Tiefgehende diskursive Erkenntnisse?
- Ressourcen
 - Zeit
 - Geld
 - Hardware
 - Verfügbarkeit von Versuchspersonen (Zielgruppe!)
- Fortschritt der Gestaltung
 - Ideenphase
 - Lo-fi Prototypen
 - Tests am realen System

Anzahl Versuchspersonen

- Das berühmte **n**
- Abhängig von Ziel und Methodik
 - Aufzeigen bestehender Usability-Probleme: mit $n \leq 15$ möglicherweise schon gut bedient
 - Statistische Hypothesentests mit Aussagen z.B. über die Bevölkerung Deutschlands: erfordert vermutlich $n > 1000$



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

Evaluationsdesign Ablauf zusammengefasst

Evaluationsziel

führt (begründet) zu

Methodenauswahl

führt (begründet) zu

Ablaufplan

Empfehlungen zum Umgang mit Versuchspersonen

- Behandeln Sie Ihre Versuchspersonen mit Respekt.
- Setzen Sie sich als Ziel, auch die Evaluation „nutzerzentriert“ zu gestalten, nicht nur das System.
- Klären Sie die Versuchspersonen vorher über den Ablauf der Evaluation auf.
- Evaluiert wird das System, nicht die Versuchsperson.

Weiterführendes: siehe *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

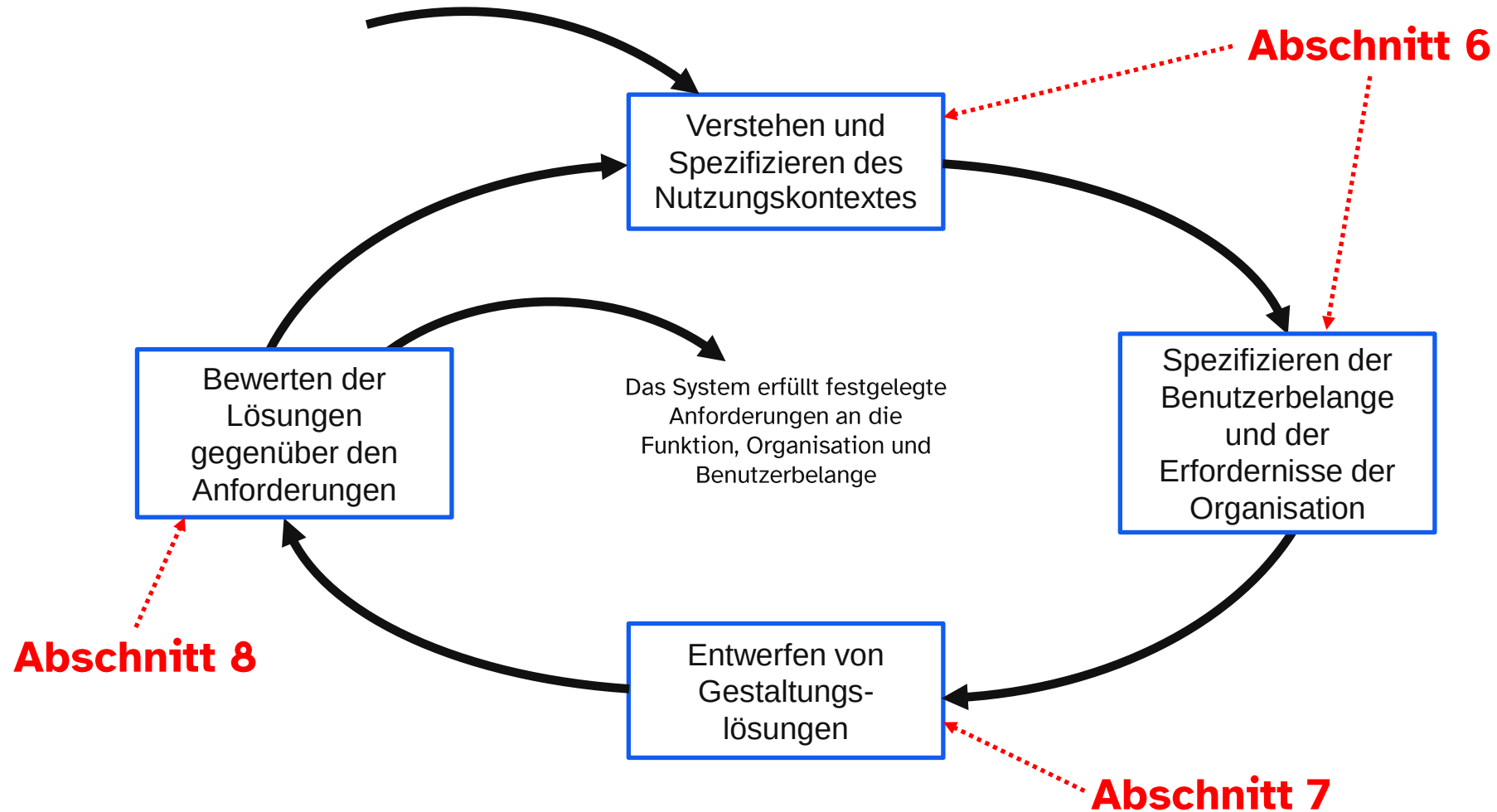
Basics and principles of empirical user studies

Hypothesis Testing

Testing hypothesis for statistical significance

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

Zur Erinnerung: Ablauf User-Centered Design



Übersicht

1	Einleitung und Überblick
2	Grundlagen der Kognition
3	Wahrnehmung und Kommunikation
4	Richtlinien für Interaktionsdesign
5	Usability Engineering Vorgehensmodelle
6	Kontextanalyse
7	Design und Prototyping
8	Evaluation
9	Interaktionsparadigmen
10	Computer-Supported Cooperative Work
11	Barrierefreiheit
12	Informationsarchitektur und Datenvisualisierung
13	Visuelle Gestaltung
14	MCI im Wandel der Zeit
15	Berufsbilder und Ethik



Teil II: Methoden

Was tun wir? In welcher Reihenfolge? Erfolgskriterien?



Rückblick: Themen dieses Abschnitts

Evaluationsmethoden: Fragebögen, Interviews, Beobachtungsstudien, Fokusgruppen, Nutzungsdaten-Logging, Eye-Tracking, Usability-Tests, Grounded Theory, Heuristische Evaluation

Evaluationsplanung: Was kann Evaluation leisten, Evaluationsziele, Planungsvorgehen, Methodenwahl

Fragen zur selbständigen Lernkontrolle:

1. Was macht Likert-Skalen aus?
2. Wann ist ein Interview semi-strukturiert?
3. Woraus kann ein Usability-Test bestehen?
4. Was ist der Unterschied zwischen den Begriffen „quantitativ“ und „objektiv“?
5. Wie viele Teilnehmer benötigt eine Evaluation? Wovon hängt die Antwort ab?