Mittelstand 4.0
Kompetenzzentrum
Usability

Usability Tests mit Lautem Denken

Ein Kompendium zur Versuchsplanung, Datenerhebung und inhaltsanalytischen Auswertung

Manfred Thüring

Berlin, Dezember 2022

Mittelstand-
Digital

Gefördert durch:
Bundesministerium
für Wirtschaft
und Klimaschutz
aufgrund eines Beschlusses
des Deutschen Bundestages

Inhaltsverzeichnis

1 Usability, User Centered Design und Usability-Tests

Die Gebrauchstauglichkeit (englisch „Usability“) eines interaktiven Systems, wie z.B. einer App, einer Website oder einer speziellen Software, stellt mittlerweile einen zentralen Qualitätsaspekt dar. Nach der Norm ISO 9241-11:2018 versteht man unter Usability …

- „… das Ausmaß, in dem ein Produkt durch bestimmte Benutzer in einem bestimmten Nutzungskontext genutzt werden kann, um Ziele in einem bestimmten Arbeitssystem effektiv, effizient und zufriedenstellend zu erreichen“.

Will man die in der Norm geforderten Attribute der Effizienz, Effektivität und Zufriedenstellung für eine Software gewährleisten, so hat dies unmittelbare Konsequenzen für die Methodik ihrer Entwicklung. Der Ansatz des User Centered Design (UCD) trägt diesen Konsequenzen Rechnung, indem er explizit die Einbeziehung zukünftiger Nutzer*innen in den Entwicklungsprozess fordert. Ausformuliert wird dies in einem Prozessmodell, das im Teil „Menschzentrierte Gestaltung interaktiver Systeme“ der Norm ISO 9241-210 (2019) beschrieben wird (vgl. Abbildung 1).

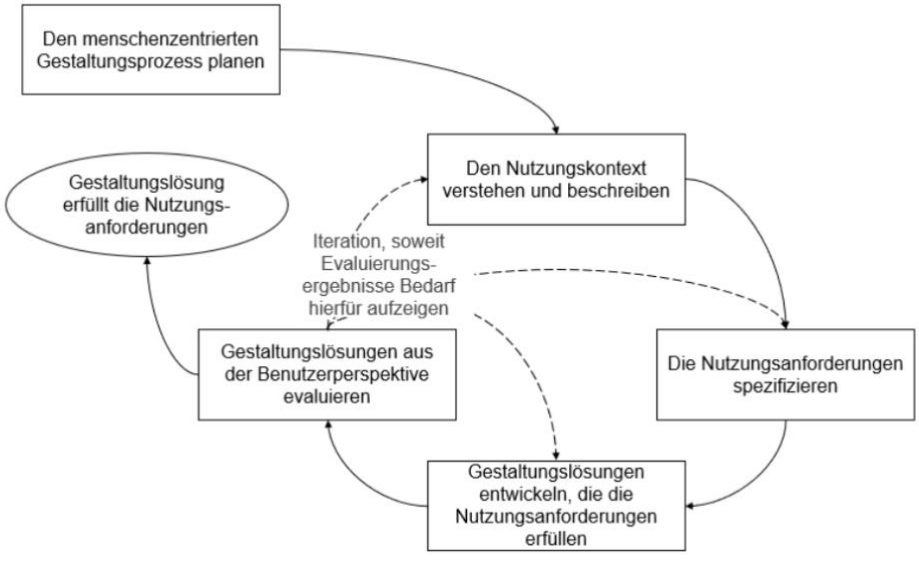

Abbildung 1: Prozess der Menschzentrierten Gestaltung interaktiver Systeme gemäß ISO 9241-210 (2019).

Ausgehend von der Planung des Gestaltungsprozesses wird zunächst eine Kontextanalyse durchgeführt, bei der Daten

- über die Verwendung des zu entwickelnden Systems (stationäre oder mobile Nutzung, Häufigkeit des Einsatzes, ...),
- über seine Nutzer*innen (Nutzerprofile, Personas, …),
- über die mit ihm zu bearbeitenden Aufgaben (Struktur, Workflow, …) sowie
- über sein zukünftiges technisches Umfeld (bestehende Infrastruktur, Schnittstellen, …)

erhoben und dokumentiert werden.

Aus den Ergebnissen der Kontextanalyse werden die Anforderungen an das System abgeleitet und sukzessive bei der Gestaltung des Systems umgesetzt. Eine Besonderheit ist dabei, dass zwischen der Entwicklung von Gestaltungslösungen und der Durchführung von Usability Tests mit potenziellen Nutzer*innen iteriert wird. Immer wenn eine neue Gestaltungsidee umgesetzt wird, wird sie evaluiert. Die dabei gewonnenen Erkenntnisse werden benutzt, um Probleme der Nutzer*innen zu erkennen und zu beseitigen. Auf diese Weise wird die Systementwicklung zyklisch vorangetrieben und solange fortgesetzt, wie die Evaluationsergebnisse hierfür Bedarf anzeigen. Treten schließlich keine Probleme mehr auf, ist die Entwicklung so weit gediehen, dass das System an den Auftraggeber übergeben bzw. dass der Markteintritt vollzogen werden kann.

Im Rahmen des UCD-Prozesses sind zwei Evaluationsformen zu unterscheiden, die als summatives und formatives Testen bezeichnet werden (Lewis, 2014).

- **Summative Usability-Tests** sollten am Ende der Systementwicklung durchgeführt werden, wenn keine weiteren Iterationen geplant sind. Ziel hierbei ist es, eine abschließende und ganzheitliche Bewertung der Usability vorzunehmen, so dass sich entscheiden lässt, ob das System tatsächlich so gebrauchstauglich ist, dass die Entwicklung beendet werden kann. Für eine solche Bewertung eignen sich vor allem standardisierte Usability-Fragebögen, die quantitative Daten in Form von Ratings produzieren, die zu einem Gesamtscore aggregiert werden

können[1]. Je höher dieser Score ist, als desto besser ist die Gebrauchstauglichkeit anzusehen.

- **Formative Usability-Tests** sollten am Ende eines jeden Entwicklungszyklus durchgeführt werden, also immer dann, wenn eine neue bzw. weiter fortgeschrittene Gestaltungslösung vorliegt. Das Ziel formativen Testens besteht darin, Usability-Probleme aus der Sicht potenzieller Nutzer*innen aufzudecken und Verbesserungen anzuregen, die im nächsten Zyklus der Iteration umgesetzt werden können. Durch dieses evolutionär ausgerichtete Vorgehen wird das zu entwickelnde System sukzessive so lange optimiert, bis keine (oder zumindest keine gravierenden) Nutzungsprobleme mehr im Usability-Test auftreten.

Der Zusammenhang zwischen dem Ausmaß der Usability eines Systems und Problemen bei seiner Nutzung wurde bereits zu einer Zeit erkannt, als der Begriff „Usability“ noch nicht gebräuchlich war und man stattdessen von „ease of use“ sprach:

- „Although it is not easy to measure „ease of use”, it is easy to measure difficulties that people have in using something. Difficulties and errors can be identified, classified, counted, and measured. So my premise is that ease of use is inversely proportional to the number and severity of difficulties people have in using software.” (Chapanis, 1981, zitiert nach Lewis, 2014, p. 665).

Demzufolge liegt es im Rahmen des UCD nahe, die formative Evaluation von Gestaltungslösungen auf die Entdeckung von Usability-Problemen auszurichten. Dies führt unmittelbar zu der Frage, was Usability-Probleme sind und wie man Usability-Tests durchführen sollte, um solche Probleme zu erkennen und zu bewerten.

[1] Beispiele sind die “Subjective Usability Scale” (SUS; Brooke, 1995) und das “Subjective Usability Measurement Inventory” (SUMI; Kirakowski, & Corbett, 1993) sowie der Fragebogen IsoNorm (Prümper, 1997) und seine Kurzversion $\text{IsoNorm}_{\text{Short}}$ (Pataki, Sachse, Prümper & Thüring, 2006).

2 Was sind Usability-Probleme – und wie lassen sie sich ermitteln?

Der Begriff Usability-Problem - ebenso wie Usability selbst – bezeichnet ein theoretisches Konstrukt[2]. „A construct is a conceptual term intended to describe a real phenomenon of theoretical interest that cannot be observed directly …" (Tractinsky, 2017, p.9). Die fehlende Möglichkeit, ein Konstrukt direkt zu beobachten, macht es erforderlich, das Konstrukt zu operationalisieren, d.h. es muss festgelegt werden, mit welchen Maßen sich seine Ausprägungen bestimmen lassen. Eine Operationalisierung gilt als gelungen, wenn die Ausprägungen des Konstrukts und die Ausprägungen der ihm zugeordneten Maße kovariieren. Da ein derartiges Messmodell die Ausprägung des Konstrukts widerspiegelt, wird es als reflektiv bezeichnet (Tractinsky, 2017).

Für das Konstrukt „Usability-Problem" gilt es Indikatoren zu spezifizieren, über die das Vorhandensein und die Schwere eines solchen Problems erschlossen werden können. Das zeigt auch die folgende Definition:

- "Ein Usability-Problem liegt vor, wenn Aspekte eines Systems es Nutzer*innen mit hinreichender Domänenerfahrung *unangenehm, ineffizient, beschwerlich oder unmöglich* machen, in einem typischen Anwendungskontext die Ziele zu erreichen, für deren Erreichung das System erstellt wurde." (Sarodnick & Brau, 2011; S. 26).[3]

Geeignete Indikatoren für unangenehme, ineffiziente, beschwerliche oder sogar vergebliche Bemühungen von Proband*innen, ihre Ziele zu erreichen, können vor allem verbale Äußerungen und beobachtete Verhaltensweisen

[2] Ob „Usability" ein wohldefiniertes und angemessen operationalisiertes Konstrukt darstellt, wird von einigen Wissenschaftlern in Frage gestellt. So charakterisiert z.B. Tractinsky (2017) Usability als ein „umbrella construct", das sich aus unscharfen Subkonstrukten zusammensetzt und nur über ein formatives Messmodell operationalisiert werden kann. Formativ meint dabei zum einen, dass die Usability Aspekte Effektivität, Effizienz und Zufriedenstellung das Konstrukt eher formen, anstatt es messbar zu machen, und zum anderen, dass keine Einigkeit über geeignete Usability-Maße besteht. Unterstützt wird diese Position u.a. durch eine Analyse von 180 Studien (Hornbaek, 2006), die mehrere Dutzend Maße für das Konstrukt und seine Subkonstrukte aufzeigte.

[3] Nutzer*innen mit Domänenerfahrung besitzen die Sachkenntnisse zur Lösung von Aufgaben, für deren Bearbeitung das zu testende System gedacht ist.

sein, die auf Schwierigkeiten bei der Interaktion mit dem technischen System schließen lassen. Werden Probandenäußerungen erfasst und Verhaltensauffälligkeiten sprachlich beschrieben, so entstehen qualitative Daten, die Rückschlüsse auf Usability-Probleme ermöglichen.

Damit diese Rückschlüsse zutreffend sind, sollten formative Usability-Tests - ebenso wie Usability-Fragebögen - den drei zentralen, psychometrischen Gütekriterien der Objektivität, Reliabilität und Validität genügen. Diese Kriterien werden folgendermaßen definiert:

- **Objektivität:** Eine Messung gilt als objektiv, wenn sie unabhängig von den Personen ist, die sie durchführen (Durchführungsobjektivität) und auswerten (Auswertungsobjektivität), sowie die erzielten Ergebnisse interpretieren (Interpretationsobjektivität).
- **Reliabilität:** Eine Datenerhebung ist zuverlässig bzw. messgenau (reliabel), wenn sie bei einer Wiederholung unter unveränderten Umständen zu denselben Ergebnissen führt.
- **Validität:** Ein valides Messinstrument misst das Konstrukt, für dessen Messung es gedacht ist.

Die drei Gütekriterien sind nicht unabhängig, sondern bauen aufeinander auf. Reliabilität setzt Objektivität, und Validität setzt Reliabilität voraus. Das Umgekehrte hingegen gilt nicht.

Inwieweit verbale Daten, wie Äußerungen der Proband*innen während eines Tests oder Interviews, diesen Kriterien genügen können, ist allerdings umstritten (Mayring, 2015). Dies gilt insbesondere für die Reliabilität und Validität formativer Usability-Tests. Positiv beeinflussen lässt sich aber deren Objektivität, wenn die Vorgehensweise bei ihrer Durchführung, Auswertung und Interpretation so weit wie möglich standardisiert ist und die Durchführenden entsprechend geschult und instruiert sind. Ist dies der Fall, sollte sich eine ausreichend hohe Reliabilität erzeugen lassen, die ihrerseits zur Validität der Messung beiträgt. Wie valide diese Messung letztendlich ist, lässt sich für qualitative Daten allerdings nur abschätzen und nicht mit denselben statistischen Methoden ermitteln, die bei quantitativen Daten angewendet werden können (vgl. auch Sedlmeier & Renkewitz, 2008). In Kapitel 8 wird diese Thematik nochmals aufgegriffen und vertieft.

3 Ziele und Fragestellungen des Tests

Unter Praktikern gilt die Methode des sog. Lauten Denkens mittlerweile als Königsweg zur Ermittlung von Usability-Problemen. Dabei werden Testpersonen instruiert laut zu denken, während sie mit dem zu untersuchenden System interagieren, um Aufgaben zu lösen, für deren Bearbeitung das System gedacht ist.

Dieses Vorgehen verfolgt zwei **Ziele**:

- Ermittlung und Beschreibung von Usability-Problemen aus Sicht der Nutzer*innen und
- Aufzeigen von Verbesserungsmöglichkeiten.

Mit den beiden Zielen sind vier zentrale **Fragestellungen** verbunden:

- Welche Funktionen und Interface-Elemente des Systems sind von Usability-Problemen betroffen?
- Wie zeigen sich Usability-Probleme in Äußerungen oder Verhaltensweisen der Nutzer*innen?
- Wie gravierend sind die auftretenden Usability-Probleme, d.h. wie ist ihr Schweregrad einzuschätzen?
- Durch welche Veränderungen am System könnten die Usability-Probleme beseitigt oder vermindert werden?

Bei der Beantwortung dieser Fragen im Rahmen des formativen Testens sind vier Schritte zu unterscheiden: Planung, Durchführung, Datenanalyse und -interpretation sowie Dokumentation der Ergebnisse.

4 Testplanung

Bei der Versuchsplanung werden die Messverfahren spezifiziert, die Prüfaufgaben und die Instruktion für die Probandinnen formuliert sowie die Art und die Größe der Stichprobe festgelegt. Dabei kann teilweise auf Ergebnisse der Kontextanalyse zurückgegriffen werden, wie z.B. auf ermittelte Workflows für die Gestaltung von Prüfaufgaben oder auf Nutzerprofile und Personas für die Auswahl von Testpersonen.

4.1 Auswahl des Messverfahrens

Zur Erhebung der Aussagen der Testpersonen während der Interaktion mit dem System können alternativ zwei Messverfahren des Lauten Denkens eingesetzt werden:

- simultanes Lautes Denken oder
- retrospektives Lautes Denken[4].

Beide Testverfahren sind sich sehr ähnlich, unterscheiden sich aber im Versuchsablauf und kommen unter verschiedenen Rahmenbedingungen zum Einsatz. Während das simultane Laute Denken die Interaktion der Versuchspersonen mit dem System begleitet, findet das retrospektive Laute Denken danach statt, d.h., die Aufgabenbearbeitung wird per Video aufgezeichnet und den Proband*innen mit der Bitte präsentiert, ihre Aktionen im Nachhinein zu kommentieren. In beiden Fällen ist die Durchführung des Tests - mit Zustimmung der Proband*in - aufzuzeichnen, da eine Videoaufnahme für die spätere Datenanalyse benötigt wird.

Findet das laute Denken simultan mit der Aufgabenbearbeitung statt, so hat dies den Vorteil, dass Äußerungen unmittelbar beim Auftreten eines Problems erfasst werden. Da sie sich noch im Kurzzeitgedächtnis befinden, können sie direkt abgerufen werden und sind weitgehend unverfälscht (Ericsson & Simon, 1980). Allerdings kann die simultane Form nicht eingesetzt werden, wenn das zu untersuchende System sprachgesteuert ist, da nicht gleichzeitig laut gedacht und per Sprache mit dem System interagiert werden kann. Von Nachteil ist weiterhin, dass die simultane Form einen Teil der Aufmerksamkeit der Versuchsperson beansprucht, was bei der Evaluation komplexer Systeme und der Untersuchung schwieriger Aufgaben zu Verzerrungen führen und die Aufgabenbearbeitung verlangsamen oder fehleranfälliger machen kann. Deshalb eignet sich das simultane Laute Denken nicht, wenn zusätzlich zu den qualitativen Daten quantitative Daten, wie Bearbeitungszeiten, Fehler oder auch Eye Tracking Parameter erfasst werden sollen.

Liegen die beschriebenen Einschränkungen vor, so sollte auf die retrospektive Form zurückgegriffen werden. Allerdings muss man sich dann darauf einstellen, dass sich Durchführungs- und Auswertungszeiten des Tests

[4] Dieses Vorgehen wird auch häufig als retrospektive Videokonfrontation bezeichnet.

mindestens verdoppeln. Für die Durchführung des retrospektiven Lauten Denkens kommen zwei Varianten in Betracht (vgl. Abbildung 4 und 5):

- **Variante A:** Aufgabenbearbeitung und Konfrontation werden nacheinander in zwei getrennten Blöcken durchgeführt.
- **Variante B:** Aufgabenbearbeitung und Konfrontation werden von den Testpersonen alternierend durchlaufen.

Welche Form des Lauten Denkens und ggf. welche Variante der retrospektiven Videokonfrontation gewählt werden sollte, hängt von der genauen Fragestellung und Zielsetzung des Tests ab. Dabei ist zu berücksichtigen, ob

- das zu untersuchende System sprachgesteuert ist,
- zusätzlich zum Lauten Danken quantitative Daten erhoben werden und / oder
- die mentale Belastung durch hohe Systemkomplexität bzw. Aufgabenschwierigkeit als groß einzuschätzen ist.

Eine Empfehlung für die Wahl von Messmethode und Varianten gibt der Entscheidungsbaum in Abbildung 2.

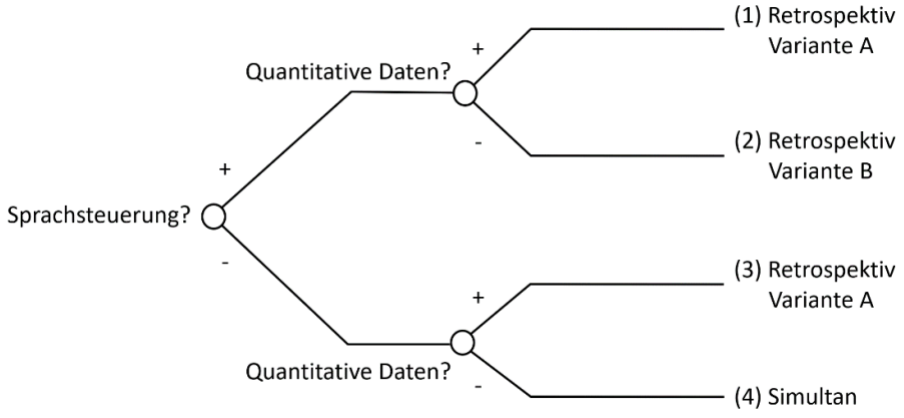

Abbildung 2: Entscheidungshilfe zur Auswahl der Messmethodik.

Die Abbildung verdeutlicht **vier Entscheidungswege**:

- Ist das System **sprachgesteuert**, so kommt nur die **retrospektive Form** Lauten Denkens in Frage.
    - Werden zusätzlich **quantitative Daten** erhoben, empfiehlt sich **Variante A** (Option 1). Andernfalls weiß die Versuchsperson

bereits nach der ersten Aufgabe, dass sie nach der Bearbeitung laut Denken soll. Dieses Wissen kann einen Einfluss auf ihre Aufmerksamkeit und ihre Leistung bei der Aufgabenbewältigung haben, wobei dieser Einfluss allerdings geringer ausfallen sollte als beim simultanen Lauten Denken.

    - Werden **keine** zusätzlichen quantitativen Daten erhoben, empfiehlt sich **Variante B** (Option 2), weil sie Gedächtniseffekte und nachträgliche Rationalisierungen durch die größere zeitliche Nähe zur Aufgabenbearbeitung verringert (Ericsson & Simon, 1980).

- Ist das System **nicht sprachgesteuert**, so kommen prinzipiell beide Messmethoden in Betracht.
    - Die **retrospektive Form** ist zu wählen, wenn quantitative Maße (wie z.B. Bearbeitungszeiten oder Häufigkeiten) nicht durch die mentale Belastung, die das Laute Denken erzeugt, verzerrt werden sollen. Damit dies völlig ausgeschlossen wird, ist **Variante A** zu empfehlen (Option 3).
    - Wird auf quantitative Daten verzichtet, ist die **simultane Form** des Lauten Denkens wegen ihrer Zeitgleichheit mit der Aufgabenbearbeitung und der damit verbundenen Authentizität der Äußerungen zu bevorzugen (Option 4).

Beide Messverfahren, ebenso wie die zwei Varianten, können sowohl im Labor als auch „remote“ mit Hilfe entsprechender Tools durchgeführt werden. In beiden Fällen, empfiehlt es sich, dass die Versuchsleitung Verhaltensauffälligkeiten, die während der Aufgabenbearbeitung auftreten, notiert. Solche Auffälligkeiten können z.B. ein Innehalten oder Zögern bei der Bearbeitung sein oder sich nonverbal, z.B. durch Kopfschütteln, Stirnrunzeln, Lächeln usw., äußern. Beim simultanen Ansatz sollten derartige Auffälligkeiten nach Abschluss des Versuchs in einem Interview angesprochen werden, beim retrospektiven Ansatz lassen sich entsprechende Fragen während der Videokonfrontation an geeigneter Stelle einflechten.

4.2 Prüfaufgaben

Gemäß der Norm ISO 9241-11:2018 spielt für die Bestimmung der Usability eines Systems seine zielorientierte Nutzung eine zentrale Rolle. Dies wird beim

formativen Testen dadurch erreicht, dass Proband*innen das zu evaluierende System zur Bearbeitung von Aufgaben, für die das System gedacht ist, einsetzen. Entsprechend ist die angemessene Auswahl und Gestaltung von Prüfaufgaben zentral für die Aussagefähigkeit eines Usability-Test.

Die Auswahl sollte sich an zwei Kriterien orientieren. Zum einen sollten in den Prüfaufgaben Arbeitsabläufe (Workflows) umgesetzt werden[5], die für die Arbeit mit dem System besonders wichtig, schwierig, komplex und/oder häufig sind – je nach spezifischer Fragestellung der Evaluation. Zum zweiten sollte die Evaluation so umfassend wie möglich sein. Wenn es nicht möglich ist, bei der Evaluation alle Funktionen und Interface-Elemente des Systems zu berücksichtigen, so sollten zumindest die wichtigsten davon für die Bearbeitung der Aufgaben benötigt werden.

Für die Gestaltung der Aufgaben ist darauf zu achten, dass sie:

- prägnant und verständlich (ggf. in der Sprache / Terminologie der Nutzer*innen) formuliert sind,
- sich in einer vertretbaren Zeit erledigen lassen[6] und
- ein klares Ziel und damit auch einen deutlich erkennbaren Abschluss haben.

Vor dem Einsatz im Test sollte der Lösungsweg jeder Aufgabe definiert werden, um ihre Machbarkeit sicherzustellen. Der Lösungsweg der Aufgabe kann außerdem bei der Datenanalyse benutzt werden, um Abweichungen von einer idealtypischen Vorgehensweise zu erkennen und dadurch Rückschlüsse auf potenzielle Usability-Probleme zu ermöglichen.

Es empfiehlt sich, die einzusetzenden Prüfaufgaben in einer Liste zusammenzustellen, die eine Kennung der Aufgabe, einen Kurztitel und den Aufgabentext enthält. Ein Beispiel dafür zeigt Tabelle 1.

---

[5] Wenn bei der Systementwicklung gemäß UCD vorgegangen wird, sollten die zentralen Workflows in der Kontextanalyse vor der ersten Iteration analysiert und dokumentiert worden sein. Darauf kann bei der Auswahl und Gestaltung der Prüfaufgaben zurückgegriffen werden.

[6] Die Dauer der Evaluation sollte möglichst 1,5 Stunden nicht überschreiten. Ist dies aufgrund des Systemumfangs nicht möglich, sollte sie auf mehrere Sitzungen verteilt werden.

Tabelle 1: Beispiel für die Gestaltung einer Prüfaufgabenliste.

Aufgabenkennung	Kurztitel	Aufgabentext
…		
Aq	Semesterstart	Ermitteln Sie auf der Website der TU Berlin, wann die Lehre im Sommersemester 2023 beginnt und nennen Sie das Datum.
Ar		
…		

Zur Vermeidung von Reihenfolgeeffekten werden die Prüfaufgaben den Proband*innen in randomisierter Reihenfolge vorgelegt. Eine Ausnahme davon besteht, wenn mehrere Aufgaben in einem sachlogischen Zusammenhang stehen. In diesem Fall sollten sie zu einem Prüfszenario mit fester Aufgabensequenz entsprechend der Sachlogik zusammengefasst werden. Die Reihenfolge der Szenarien ist dann ihrerseits zu randomisieren.

4.3 Weitere Versuchsmaterialien

Materialien, die für die Planung und Umsetzung eines Usability-Tests benötigt werden, sind:

- ein kurzer Begrüßungstext,
- eine Aufklärung über den Umgang mit persönlichen Daten in Kombination mit einer Einverständniserklärung und einer Anleitung zur Erzeugung einer anonymen Versuchspersonenkennung,
- eine Instruktion, die den Versuchspersonen erläutert, was ihre Aufgabe ist und wie der Test abläuft sowie
- Fragen zur Demografie.

Beispiele hierzu sind im Anhang 1 zur Datenerhebung zu finden.

4.4 Stichprobe: Zusammensetzung und Umfang

Bei der Bildung der Stichprobe für einen formativen Usability-Test ist zu entscheiden, welche Personen am Test teilnehmen und wieviele Proband*innen benötigt werden.

Ein **Kardinalfehler**, der häufig bei der Stichprobengestaltung (meist aus pragmatischen Gründen) gemacht wird, besteht darin, leicht verfügbare Personen oder sogar die Entwickler*innen des Systems selber als Testpersonen

einzusetzen. Beide Personengruppen haben in der Regel nicht das nötige Domänenwissen, um die Sachprobleme, die die Prüfaufgaben beinhalten, effektiv und effizient zu lösen. Für Entwickler*innen kommt erschwerend hinzu, dass sich für sie kaum Interaktionsprobleme ergeben, da sie mit dem System bestens vertraut sind[7]. Für beide Personengruppen ist deshalb mit einer erheblichen Verzerrung der Testergebnisse zu rechnen. Diese Verfälschung der Daten lässt sich vermeiden, wenn Personen aus dem Kreis der zukünftigen Nutzer*innen des Systems als Versuchspersonen gewonnen werden. Die wichtigste Regel bei der Bildung der Stichprobe lautet also: „Setze nur potenzielle Nutzer*innen für den Test ein!“

Als nächstes stellt sich die Frage, wieviel Nutzer*innen für den Test benötigt werden. Für eine Antwort darauf haben Nielsen und Landauer (1993) ein mathematisches Modell entwickelt. Setzt ein Usability-Test i Versuchspersonen[8] ein, so ergibt sich die Anzahl gefundener Usability-Probleme aus der Formel

$$\text{Found}(i) = N\,(1 - (1 - L)^i).$$

Hierbei bezeichnet Found (i), die Anzahl der von i Testpersonen *insgesamt entdeckten* Probleme, N die Anzahl der *insgesamt bestehenden* Usability-Probleme, i die Stichprobengröße und L die Wahrscheinlichkeit für die Entdeckung eines *einzelnen Problems* durch *eine einzelne Testperson*. In einer Reihe von Studien zeigte Nielsen (2000), dass 0,31 ein typischer Wert für die Entdeckungswahrscheinlichkeit L ist. Setzt man diesen Wert in Bezug zur Anzahl der eingesetzten Versuchspersonen, so erhält man die Schätzung, dass mit einer Stichprobengröße von 15 Personen 100% der Usability-Probleme entdeckt werden und dass ca. 5 Personen genügen, um 85% der Probleme zu identifizieren.

[7] Die Unterscheidung von Sach- und Interaktionsproblem geht zurück auf Streitz (1985, 1986).

[8] Abweichend von der hier praktizierten Schreibweise wird die Stichprobengröße üblicherweise nicht mit i sondern mit n bezeichnet.

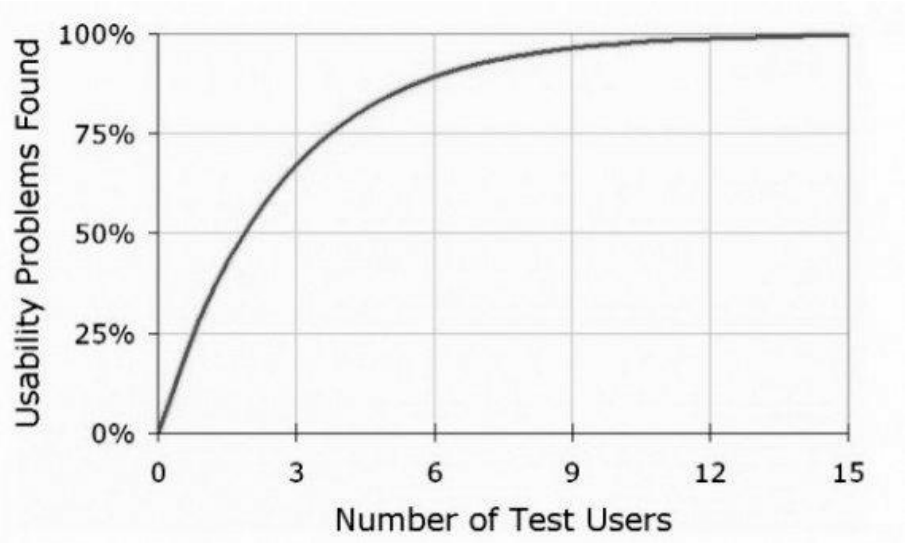

Abbildung 3: Schätzung der entdeckten Usability-Probleme in Abhängigkeit von der Anzahl der Versuchspersonen bei einer Entdeckungswahrscheinlichkeit von 0,31. Vergleiche: https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/

Von dem Maximalwert 15 ausgehend, stellt Nielsen eine Kosten-Nutzen-Berechnung an, nach der sich bei einem Einsatz von nur 3 Versuchspersonen das beste Kostennutzenverhältnis für einen Usability-Test erzielen lässt (vgl. Abbildung 3), und empfiehlt aus Gründen der Testeffizienz einen Richtwert von 5 Testpersonen. Dabei argumentiert er, dass (unter der Voraussetzung eines ausreichenden Testbudgets für 15 Proband*innen) lieber 3 mal 5 Personen als 1 mal 15 Personen das System evaluieren sollten. Unterstützt wird Nielsens Schätzung durch Ergebnisse von Virzi (1992), der in einer Studie zeigte, dass mit n=5 ca. 80% der Usability-Probleme entdeckt werden können.

An Nielsens Empfehlung einer Größe von n=5 ist allerdings von verschiedenen Seiten Kritik geübt worden. So fanden in einem Usability-Test von Perfetti (2001) für eine e-commerce Webseite die ersten 5 Proband*innen einer Stichprobe mit n=18 lediglich 35% der Probleme. Betrachtete man die gesamte Stichprobe, so zeigte sich außerdem, dass (a) von jeder hinzukommenden Person mehr als 5 neue Usability-Probleme detektiert wurden, (b) sich viele der ernsten Probleme erst in den späten Tests zeigten und dass (c) die Rate der mehrfach entdeckten Probleme im Verlauf der Tests nicht wie von Nielsen postuliert anstieg.

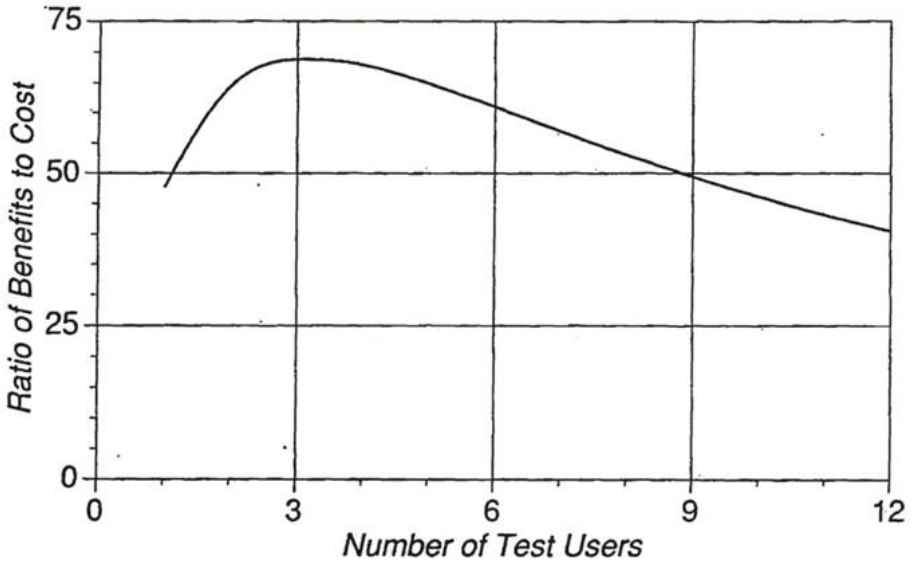

Abbildung 4: Kosten-Nutzen-Verhältnis der Problemdetektion in Abhängigkeit von der Anzahl der Versuchspersonen (vereinfachte Abbildung aus Nielsen und Landauer (1993, S. 212).

Faulkner (2003) unterteilte in einer Untersuchung 60 Nutzer*innen in Gruppen verschiedener Größe. Einige Gruppen von n=5 entdeckten 99% der Probleme, andere hingegen nur 55%. Bei einer Größe von n=10 wurden im Minimum 80% der Probleme, im Maximum 95% der Probleme detektiert. Die Ergebnisse dieser und weiterer Studien (vgl. hierzu Lewis, 2014) stehen in deutlichem Widerspruch zu Nielsens Schätzung.

Lewis (2014) stellt fest, dass eine Stichprobe von 5 Personen durchaus angemessen sein kann, dies aber nur unter sehr eingeschränkten Bedingungen. Denn wenn das zu untersuchende System eine hohe Zahl von Usability-Problemen aufweist und der prozentuale Anteil der von nur einer Person entdeckten Probleme im Durchschnitt weniger als 31% beträgt, dann gibt es keine Garantie, dass 5 Personen für eine Entdeckungsrate von 85% ausreichen. Lewis empfiehlt deshalb, sich nicht blind auf die „magical number 5“ zu verlassen, sondern jedes Mal genau abzuwägen, welches n zu wählen ist.

Hierfür stellt er selber Berechnungen analog zu Nielsen und Landauer (1993) an, und fasst die Ergebnisse tabellarisch zusammen (vgl. Tabelle 2)[9].

Tabelle 2: Wahrscheinlichkeiten der Entdeckung von Usability-Problemen in Abhängigkeit von Stichprobengröße n und Auftretenswahrscheinlichkeit p (Lewis, 2014, S.678).

p	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=10$	$n=15$	$n=20$
.01	0.01	0.02	0.03	0.04	0.05	0.10	0.14	0.18
.05	0.05	0.10	0.14	0.19	0.23	0.40	0.54	0.64
.10	0.10	0.19	0.27	0.34	0.41	0.65	0.79	0.88
.15	0.15	0.28	0.39	0.48	0.56	0.80	0.91	0.96
.25	0.25	0.44	0.58	0.68	0.76	0.94	0.99	1.00
.50	0.50	0.75	0.88	0.94	0.97	1.00	1.00	1.00
.90	0.90	0.99	1.00	1.00	1.00	1.00	1.00	1.00

In Tabelle 2 bezeichnet n die Stichprobengröße und p die Wahrscheinlichkeit dafür, dass ein Problem bei einer Versuchsperson auftritt und somit mindestens einmal entdeckt wird. Führt man einen Usability-Test mit 5 Personen durch, dann beträgt nach Tabelle 2 die Wahrscheinlichkeit dafür, dass alle Usability-Probleme bei einer Auftretenswahrscheinlichkeit von 0.50 mindestens einmal entdeckt werden, 0.97. Tabelle 2 zeigt auch, dass die Anzahl n der benötigten Probanden umso größer wird, je geringer p und je größer die angestrebte Entdeckungsraten ausfallen.

Eine genaue Abschätzung der erforderlichen Stichprobengröße gestaltet sich in der Praxis allerdings als schwierig, da die Auftretenswahrscheinlichkeit p im Vorfeld nicht bestimmbar ist, sondern hierzu lediglich nur Annahmen getroffen werden können. Lewis (2014) betont deshalb, dass seine Berechnungen vor allem den Zweck verfolgen, realistische Erfolgsabschätzungen zu ermöglichen, und dazu anregen sollen, sich im Vorfeld über die genaue Zielsetzung des Tests Gedanken zu machen. Zusammenfassend erklärt er: "… a study with n=5 is likely to leave many problems undiscovered if their probability of occurrence is less than .50, but, more optimistically, the study will probably uncover

[9] Lewis verwendet als Basis seiner Berechnung die kumulative Binomialverteilung und leitet für die Stichprobengröße folgende Formel ab: n = ln(1 – discovery goal / ln(1-p). Das discovery goal entspricht in Tabelle 3 den jeweiligen Werten in den Zellen.

enough problems to give the developers the information needed to improve the product's usability." (Lewis, 2014, S. 678). Dies gilt vor allem dann, wenn man im Rahmen des UCD-Prozesses formativ vorgeht und entsprechend mehrfach, also am Ende eines jeden Entwicklungszyklus, testet.

Da sich mit n = 15 nach Lewis schon bei einer Auftretenswahrscheinlichkeit von nur .15 eine Detektionsrate von 91% erzielen lässt, erscheint diese Stichprobengröße als Obergrenze[10] aus praxisorientierter Perspektive als angemessen. Steht nur ein geringes Budget zur Verfügung, erzielt man aber auch schon mit n = 3 einen bedeutsamen Erkenntnisgewinn, so dass die gewählte Stichprobengröße in der Regel zwischen den Grenzen 3 und 15 liegen sollte.

Neben der Spezifikation der angestrebten Detektionsrate spielt für die insgesamt zu wählende Anzahl von Testpersonen noch eine weitere Überlegung eine Rolle. Bevan, Kirakowski und Maissel (1991) charakterisieren Usabilty als „Quality of Use": „A product is not itself usable or unusable, but has attributes which will determine the usability for a particular user, task and environment." (Bevan et al., 1991, S.4; zitiert nach Tullis, 2019). Folgt man dieser Definition, so kommt es insbesondere bei komplexeren Softwarepaketen, wie z.B. CRM- oder ERP-Systemen[11], häufig vor, dass verschiedene Funktionen bzw. Module des Systems von Personengruppen verwendet werden, die sich hinsichtlich Expertise, Aufgaben und Arbeitsumfeld voneinander unterscheiden. Will man diese Heterogenität bei der Evaluation berücksichtigen, so werden mehrere Stichproben benötigt, die im Test unterschiedliche Systembestandteile einsetzen, um unterschiedliche Aufgaben zu bearbeiten. Hierfür erweist sich ein Rückgriff auf Nutzerprofile und Personas, die in der Kontextanalyse des UCD-Prozesses erstellt wurden, als hilfreich. Für die Größe der einzelnen Stichproben gelten dann die oben beschriebenen Ansätze.

[10] Ausnahmen hiervon stellen Systeme dar, für die die Detektionsrate besonders hoch sein sollte, wie z.B. bei sicherheitskritischen Systemen, da software-ergonomische Mängel oft eine Mitschuld an Unfällen tragen (Leveson, 1995).

[11] CRM ist die Abkürzung für Customer Relation Management, ERP für Enterprise Ressource Planning.

4.5 Versuchsdesign und Versuchsablauf

Für das Versuchsdesign sind **drei Fälle** zu unterscheiden:

1. Im einfachsten Fall evaluieren alle Testpersonen dieselben Systembestandteile (Funktionen, Module) mit denselben Prüfaufgaben bzw. -szenarien in randomisierter Reihenfolge.
2. Sollen Personen mit einem bestimmten Profil getestet werden, so ist festzulegen, welche Prüfaufgaben bzw. Prüfszenarien für sie geeignet sind. Aus der Bearbeitung der Aufgaben ergibt sich, welche Funktionen / Module des Systems evaluiert werden. Die Reihenfolge der Aufgaben / Szenarien ist zu randomisieren.
3. Gilt es, mehrere Profile bei der Evaluation zu berücksichtigen, so sollte dies in getrennten Gruppen, die unabhängig voneinander arbeiten, erfolgen. Für jede Gruppe sind die jeweiligen Prüfaufgaben und deren randomisierte Folge festzulegen.

Für die Zusammensetzung der jeweiligen Stichproben und die Zuordnung der Prüfaufgaben kann auf die Ergebnisse der Kontextanalyse zurückgegriffen werden. Für eine effiziente Versuchsdurchführung und die Gewährleistung einer hohen Durchführungsobjektivität sollte das Versuchsdesign schriftlich in einem Versuchsplan dokumentiert werden, z.B. in Form einer Liste (vgl. Tabelle 3).

Aus dem Beispiel der Tabelle ist ersichtlich, dass Versuchspersonen nicht namentlich erfasst werden, sondern eine anonyme Kennung erhalten. Des Weiteren zeigt das Beispiel, dass die Reihenfolge der Prüfaufgaben für die Versuchspersonen randomisiert ist. Die Kennung der Prüfaufgaben muss mit der Kennung übereinstimmen, die in der Aufgabenliste (vgl. Tabelle 1) angegeben ist. Für den dritten der oben genannten Fälle kann für jede der Gruppen eine separate Tabelle angelegt oder die Tabelle um eine Spalte ergänzt werden, in der die Gruppenzugehörigkeit vermerkt wird.

Bei formativen Usability-Tests können sowohl qualitative als auch quantitative Daten anfallen. Qualitative Daten sind sprachliche Aussagen, die beim Lauten Denken (simultan oder retrospektiv) oder bei nachfolgenden Interviews entstehen. Diese Daten können mit inhaltsanalytischen Verfahren ausgewertet werden (vgl. Kapitel 6). Quantitative Daten in Form von Häufigkeiten entstehen, wenn sprachliche Aussagen quantifiziert werden. Bei der

retrospektiven Testvariante ist es außerdem möglich, Bearbeitungszeiten oder Fehlerhäufigkeiten zu erheben.

Tabelle 3: Beispiel für einen Versuchsplan mit n Versuchspersonen und m randomisierten Prüfaufgaben.

Versuchsperson	**Versuchsdurchgang**	**Prüfaufgabe**
00X	1	Ar
00X	2	Aq
00X	…	…
00X	m	Ax
00Z	1	Aq
…	2	Ax
n	…	…

Für den dritten der oben genannten Fälle kann für die Auswertung derartiger Daten prinzipiell ein faktorielles Versuchsdesign aufgestellt werden, in dem z.B. die Gruppen als Faktor dienen. Für die Auswertung derartiger Designs kommen spezielle inferenzstatistische Methoden in Betracht, wie z.B. t-Tests, Varianzanalysen oder verteilungsfreie Verfahren. Mit ihnen könnten z.B. die verschiedenen Gruppen verglichen werden, um Unterschiede zwischen ihnen zu ermitteln. Haben die Gruppen verschiedene Aufgaben bearbeitet und/oder unterschiedliche Systemfunktionen benutzt, ist ein solcher Vergleich allerdings wenig sinnvoll, da die Gruppenzugehörigkeit mit den Prüfaufgaben und den eingesetzten Funktionalitäten konfundiert ist, so dass ermittelte Unterschiede nicht eindeutig zu interpretieren sind. Entsprechend kann man in diesem Fall auf ein faktorielles Design verzichten und sollte sich bei der Auswertung der quantitativen Daten auf deskriptive Statistiken beschränken[12].

5 Testdurchführung

Je nach gewählter Messmethode kann das Laute Denken simultan zur Bearbeitung der Prüfaufgaben erfolgen oder retrospektiv nach Abschluss der Bearbeitung. Zur Gewährleistung einer hohen Durchführungsobjektivität sollte

[12] Folgt man den Empfehlungen zur Stichprobengröße in Kapitel 4.4, so besteht eine weitere Einschränkung darin, dass Unterschiede zwischen den Gruppen sich aufgrund der geringen Probandenzahl wahrscheinlich statistisch nicht nachweisen lassen und das erforderliche Signifikanzniveau verfehlen.

der Ablauf der Datenerhebung standardisiert erfolgen, d.h. für verschiedene Testleiter*innen und Versuchspersonen möglichst immer identisch sein. Dafür empfiehlt sich die Verwendung von Leitfäden, an denen sich die Testleiter*innen orientieren können. Dies gilt für alle vier Varianten des Lauten Denkens (simultan vs. retrospektiv und laborgestützt vs. Remote).

5.1 Simultanes Lautes Denken

Die Datenerhebung mit Lautem Denken umfasst fünf Phasen, deren einzelne Schritte sich in einem Leitfaden zusammenfassen lassen (vgl. Tabelle 4).

In der **Vorbereitungsphase** wird zunächst die Testumgebung eingerichtet sowie die benötigte Technik hochgefahren und ihre Funktionalität geprüft. Für das **Labor** bedeutet dies, dass die Temperatur, die Beleuchtung und ggf. der Schallschutz geregelt werden, so dass über alle Durchgänge hinweg und für alle Proband*innen die äußeren Bedingungen des Tests gleich sind. Dadurch werden mögliche, äußere Störeffekte minimiert. Außerdem müssen die benötigte Hardware und Software gestartet werden, wie z.B. Versuchsrechner, Kameras und Aufzeichnungsgeräte. Beim Einsatz komplexerer Technik empfiehlt sich die Verwendung einer Checkliste, an der sich die Versuchsleitung orientieren kann, um keine Einstellung zu übersehen oder zu vergessen. **Remote Tests** können am besten per Videokonferenz (z.B. mit Zoom, WebEx oder einem anderen Konferenzsystem) durchgeführt werden, so dass eine sichere und stabile Internetverbindung vorhanden sein muss.

Neben der Technik werden sowohl beim Labortest als auch beim remote Test weitere Materialien in analoger oder digitaler Form benötigt. Hierzu zählen Instruktion, Fragebögen, Prüfaufgaben und ggf. Erhebungsbögen für Notizen der Versuchsleitung sowie der Versuchsplan. Mit Hilfe des Versuchsplans wird die Reihenfolge der Prüfaufgaben bestimmt und der Versuchsperson zur Anonymisierung ihrer Daten eine Kennung zugeteilt[13]. Alle Vorbereitungen sollten abgeschlossen sein, ehe die Versuchsperson erscheint bzw. zur Videokonferenz eingeladen wird.

[13] Die Kennung kann mit einem vorgegebenen Verfahren auch von der Versuchsperson selber erzeugt werden (vgl. Anhang 1).

Tabelle 4: Leitfaden zur Datenerhebung mit Lautem Denken.

Phase	**To Do**
Vorbereitung	Einrichtung der Testumgebung und Hochfahren der Technik
	Funktionsprüfung
	Zusammenstellung der Versuchsunterlagen
	Zuweisung der Versuchsperson zum Versuchsdesign
Briefing	Begrüßung der Versuchsperson
	Kurzvorstellung des Systems (Name, Art, ...)
	Aufklärung der Versuchsperson
	Einholen einer Einverständniserklärung
Datenerhebung	Demografische Daten
	Ggf. Einweisung in das System oder kurzes Training
	(Vor-)Lesen der Instruktion
	Durchführung eines Trainingsdurchgangs
	Bearbeitung der Prüfaufgaben
	Begleitende Beobachtung und Notizen der Versuchsleitung
	Ggf. ergänzendes Interview
	Ggf. abschließender Fragebogen
Debriefing	Ggf. Bezahlung
	Ggf. weitere Erläuterungen und Verabschiedung
Nachbereitung	Überprüfung der Datensicherung und Backup
	Herunterfahren der Testumgebung

Beim **Briefing** im Labor wird zunächst die Versuchsperson begrüßt, ehe sie weitere Informationen zum Test erhält. Hierzu zählen eine Kurzvorstellung des Systems und mündliche Angaben zu Ziel, Ablauf und Durchführungszeit der Untersuchung sowie zur Speicherdauer der anonymisierten Daten. Ein sehr wichtiger Punkt betrifft die Aufklärung darüber, dass die Versuchsperson jederzeit ohne Begründung und Nachteile den Test abbrechen sowie die Löschung ihrer Daten auch schon vor Ablauf der Speicherdauer fordern kann[14]. Des Weiteren ist zu erläutern, dass der Test für die Datenanalyse aufgezeichnet werden soll und die Versuchsperson wird gebeten, eine entsprechende Einverständniserklärung zu unterschreiben (vgl. Anhang 1 für ein Beispiel).

[14] Um dies zu ermöglichen, ist es notwendig, dass die Versuchsperson die Kennung ihrer Daten kennt, z.B. dadurch, dass sie diese nach dem genannten Verfahren selbst erzeugt hat.

Findet der Test remote statt, wird die Versuchsperson per Mail eingeladen und erhält dadurch Zugang zur Videokonferenz. Das Briefing findet in diesem Fall digital statt und umfasst die gleichen Punkte wie im Labor, mit dem Unterschied, dass die Einverständniserklärung der Versuchsperson per Mail oder Chatfunktion geschickt werden muss[15]. Akzeptiert sie die Erklärung, kann die Aufzeichnung gestartet und die Person gebeten werden, die Erklärung vorzulesen und zu bestätigen. Dadurch kann auf das Unterschreiben der Erklärung verzichtet werden.

Die **Datenerhebung** beginnt in der Regel mit der Erfassung demografischer Daten anhand eines Fragebogens (vgl. Anhang 1 für ein Beispiel). Alternativ ist es aber auch möglich, dies auf das Ende des Tests zu verschieben und vor dem Debriefing durchzuführen. Je nach Ausgangssituation und Fragestellung kann zunächst das System erläutert oder auch kurz durch die Teilnehmer*innen exploriert werden. Ist man allerdings daran interessiert, ob Personen das System intuitiv, also ohne diese einleitenden Maßnahmen, verstehen, wird auf diesen Teil verzichtet. Im nächsten Schritt erhalten der Proband*innen die Instruktion (vgl. Anhang 1) mit der Bitte, diese zu lesen und ggf. Fragen jetzt zu stellen, um Unterbrechungen während der Versuchsdurchführung zu vermeiden. Die Instruktion sollte ihnen im Verlauf der Durchführung weiterhin zur Verfügung stehen, falls sie noch einmal etwas nachschlagen möchten. In der Instruktion wird die eingesetzte Methodik erläutert und betont, dass es wichtig ist, alle positiven und negativen Eindrücke zu verbalisieren, die bei der Aufgabenbearbeitung entstehen. Damit sich die Proband*innen an das ungewohnte Laute Denken gewöhnen können, empfiehlt sich ein Übungsdurchgang mit ein bis zwei Prüfaufgaben. Diese Aufgaben sollten für aller Versuchspersonen identisch sein und nur für die Übung verwendet werden.

Im Anschluss an die Übung findet die eigentliche Datenerhebung statt. Dabei werden die Prüfaufgaben den Proband*innen schriftlich in der Reihenfolge vorgelegt, die im Versuchsplan für sie festgelegt wurde. Zu Beginn jeder Bearbeitung sollte die Versuchsperson die jeweilige Aufgabe laut vorlesen, um

[15] Dies kann auch im Vorfeld bei der Probandenakquise geschehen und mit der Bitte verbunden werden, die Erklärung unterschrieben zurückzuschicken, ehe der Test durchgeführt wird.

sicherzustellen, dass sie sie einmal in Gänze verarbeitet hat. Außerdem hat sie Gelegenheit, Fragen zur Aufgabe zu stellen, damit Verständnisprobleme beseitigt werden können. Während der Aufgabenbearbeitung sind Fragen zu vermeiden, da sie das Laute Denken unterbrechen und das Verbalisieren stören würden. Ist eine Prüfaufgabe beendet, so wird zur nächsten Frage übergegangen. Kann aufgrund eines schwerwiegenden Usability-Problems eine Aufgabe nicht fertiggestellt werden kann, so wird die Bearbeitung abgebrochen, ehe der Versuchsperson die nächste Aufgabe vorgelegt wird. Ist die Versuchsperson durch das Scheitern an einer Aufgabe irritiert, sollte (wie auch schon in der Instruktion) betont werden, dass nicht die Versuchsperson, sondern das System getestet wird.

Während des Tests sollte sich die Versuchsleitung neutral verhalten. Eine Interaktion mit der Versuchsperson ist möglichst zu vermeiden und auf jegliche Bewertung der Bearbeitung zu verzichten. Dies gilt auch für das nonverbale Verhalten, wie z.B. Stirnrunzeln, Lächeln, aufmunterndes Nicken etc. Des Weiteren sollte der Versuchsperson nur dann bei der Aufgabenbearbeitung geholfen werden, wenn sie darum bittet. Zusätzlich zur Präsentation der Aufgaben muss die Versuchsleitung das Laute Denken überwachen, um sicherzustellen, dass sich die Verbalisierungen nicht verringern oder sogar ganz aufhören. Werden derartige Tendenzen bemerkt, so ist die Versuchsperson zu bitten, das Laute Denken aufrechtzuerhalten. Dies erweist sich bei längeren Versuchsdauern und zunehmender Ermüdung der Proband*innen häufig als notwendig. Während der Aufgabenbearbeitung hat die Versuchsleitung Gelegenheit, sich Auffälligkeiten (wie z.B. Bitten um Unterstützung) zu notieren.

Wird der Test im Labor durchgeführt, können Instruktion und Prüfaufgaben ausgedruckt den Probandinnen vorgelegt werden. Erfolgt der Test remote, so empfiehlt es sich, die Prüfaufgaben per Chatfunktion zu übermitteln, Instruktionen aufgrund ihrer Länge hingegen vor Beginn des Tests per Mail zu verschicken.

Sind alle Prüfaufgaben bearbeitet worden, ist die Phase des Lauten Denkens beendet. Bei Bedarf kann anschließend ein Interview durchgeführt werden. Ist dafür im Vorfeld ein Leitfaden erarbeitet worden, so sollte dieser durch gezielte Fragen auf Basis der Notizen der Versuchsleitung ergänzt werden.

Im **Debriefing** gibt es Gelegenheit, Fragen zu beantworten oder auch weitere Erläuterungen zum Test abzugeben. Wurde ein Versuchspersonenhonorar vereinbart, so wird es ausgezahlt und quittiert, ehe die Versuchsperson verabschiedet wird. Beim remote Test kann die Entlohnung per Überweisung oder durch einen Gutschein erfolgen.

Zur **Nachbereitung** gehört die Kontrolle der Datenaufzeichnung und es empfiehlt sich ein Back-up des Recordings anzulegen, ehe die Versuchsumgebung heruntergefahren wird.

5.2 Retrospektives Lautes Denken

Vorbereitung, Briefing und die ersten Schritte der Durchführung (demografische Datenerhebung, Systemeinweisung und Trainingsdurchgang) verlaufen analog zum simultanen Lauten Denken. Der wesentliche Unterschied zwischen beiden Messverfahren besteht darin, dass die Versuchspersonen bei der retrospektiven Videokonfrontation ihre Interaktion mit dem System nicht während der Aufgabenbearbeitung, sondern erst danach sprachlich begleiten. Für die Durchführung des Tests kommen die beiden in Abschnitt 4.2.1 beschriebenen Varianten in Betracht.

5.2.1 Variante A: Geblockte Durchführung

Der Ablauf unter Variante A ist in zwei getrennte, aufeinander folgende Blöcke unterteilt, denen jeweils eine eigene Instruktion vorangestellt ist (vgl. Abbildung 5).

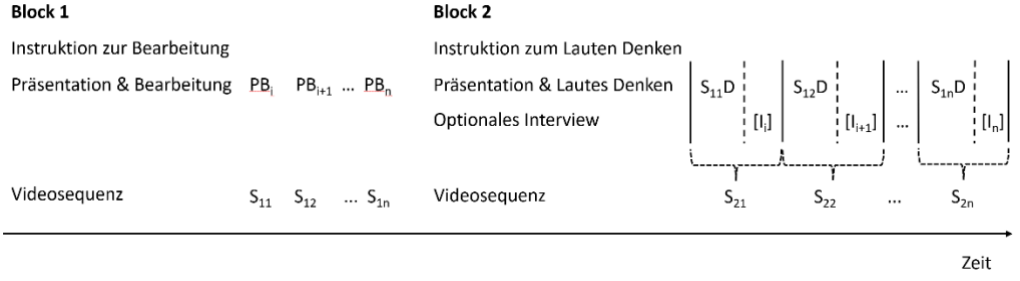

Abbildung 5: Variante A zur Durchführung des retrospektiven Lauten Denkens.

Zu Beginn des ersten Blocks wird die Testperson zunächst gebeten, die Instruktion einmal laut vorzulesen. Um die Vorteile der Variante A zu realisieren, sollte diese Instruktion auf die Aufgabenbearbeitung fokussiert

sein und das Laute Denken noch nicht erwähnen (vgl. Anhang 1 für ein Beispiel).

Nach der Instruktion werden alle Prüfaufgaben (P_i bis P_n) in der randomisierten Folge dargeboten, die für die jeweilige Testperson bei der Planung im Versuchsplan festgelegt wurde. Die Bearbeitung wird ohne Unterbrechung aufgezeichnet, lässt sich aber in einzelne Videosegmente (S_{11} bis S_{1n}) unterteilen, die für Block 1 die Lösung der jeweiligen Prüfaufgabe dokumentieren. Die Versuchsleitung beobachtet die Aufgabenbearbeitung und notiert etwaige Auffälligkeiten. Besonders wichtig für die Beobachtung sind Indikatoren, die auf Usability-Probleme hinweisen, wie z.B. das Abkommen vom kürzesten Lösungsweg einer Aufgabe, verlängerte oder gescheiterte Suchprozesse, Navigations- oder Orientierungsprobleme sowie negative Kommentare, die die Versuchsperson unaufgefordert äußert. Nach Abschluss von Block 1 wird die Aufzeichnung gespeichert.

In der Instruktion zu Block 2 (vgl. Anhang 1 für ein Beispiel) wird der Testperson mitgeteilt, dass ihr nun die Bearbeitung der Prüfaufgaben per Video gezeigt wird, verbunden mit der Bitte, sich möglichst genau an ihre Gedanken und Empfindungen während der Interaktion mit dem System zu erinnern und diese zu verbalisieren. Nach der Instruktion werden die Videosequenzen (S_{11} bis S_{1n}) in der Reihenfolge der Prüfaufgaben aus dem ersten Block dargeboten. Simultan zur Darbietung erfolgt nun das Laute Denken (D). Nach jeder Prüfaufgabe kann die Versuchsleitung gezielte Fragen auf der Grundlage ihrer Beobachtungsnotizen und ein kurzes Interview führen ([I_i] bis [I_n]). Das Laute Denken und die jeweiligen Interviews werden aufgezeichnet und ergeben die Videosequenzen für Block 2 (S_{21} bis S_{2n}).

5.2.2 Variante B: Alternierende Durchführung

Der Ablauf für die alternierende Durchführung unterscheidet sich wesentlich von dem der Variante A. In der Instruktion zu Variante B wird erläutert, dass eine Reihe von Prüfaufgaben zu bearbeiten sind und dass die Bearbeitung aufgezeichnet wird. Des Weiteren wird erklärt, dass nach der Bearbeitung die Aufzeichnung gezeigt wird, verbunden mit der Bitte, sich an die Bearbeitung zu erinnern und zu verbalisieren, was dabei gedacht und empfunden wurde.

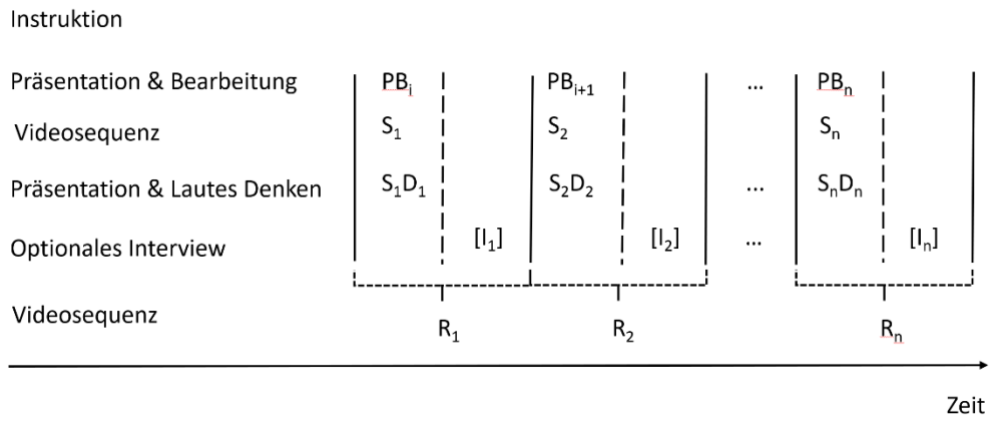

Abbildung 6: Variante B zur Durchführung des retrospektiven Lauten Denkens.

Nach der Instruktion wird die erste Prüfaufgabe gemäß Versuchsplan (PB_i) dargeboten und bearbeitet. Die Bearbeitung wird aufgezeichnet (S_1) Ist die Bearbeitung beendet, wird die erste Aufzeichnung präsentiert und die Testperson denkt dazu laut (S_1D_1). Im Anschluss hat die Testleitung die Option, auf Basis angestellter Beobachtungen ein kurzes Interview zu führen und Fragen zu stellen [I_1]). Der erste Versuchsdurchgang ist damit abgeschlossen und durch die Videosequenz R_1 dokumentiert. Dieses Vorgehen wird wiederholt, bis jede der n bearbeiteten Prüfaufgaben gezeigt und durch lautes Denken begleitet wurde (S_1D_1 bis S_nD_n). Zusammen mit den geführten Interviews ist dies in den Videosequenzen R_1 bis R_n dokumentiert.

Zum Abschluss des Tests mit der retrospektiven Messmethode erfolgen Debriefing und Nachbereitung analog zum simultanen Lauten Denken.

5.3 Aufgezeichnete audio-visuelle Daten

Nach Durchführung des Usability-Tests liegt von jeder Versuchsperson eine Bild- und Tonaufzeichnung vor. Absolvieren zum Beispiel sechs Versuchspersonen jeweils fünf Prüfaufgaben, so besteht die entstandene Materialsammlung aus sechs Videos, die zusammen 30 Interaktionsverläufe sowie alle Äußerungen und Interviews der Versuchspersonen beinhalten. Diese Sammlung von audio-visuellen Daten zeichnet sich durch eine Fülle an Informationen aus, die nicht nur die Usability-Problematik betreffen (vgl. Abb. 7).

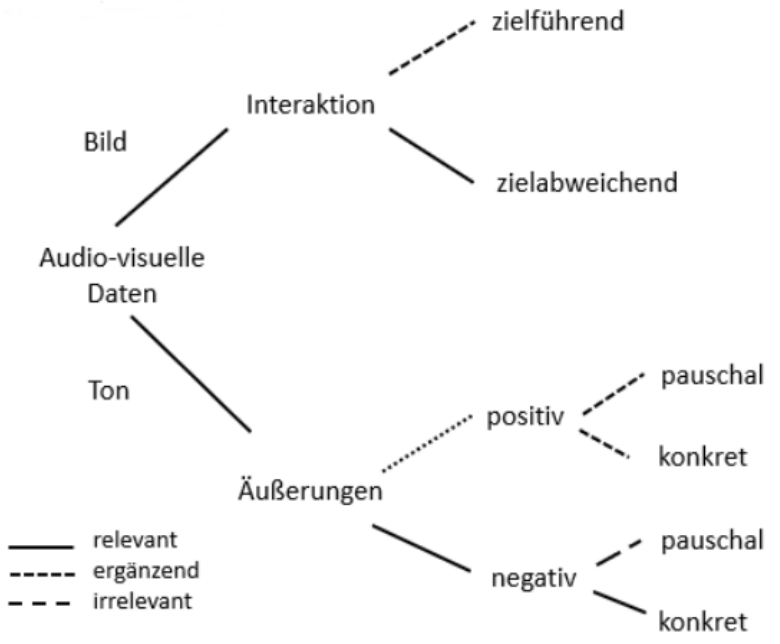

Abbildung 7: Struktur der audio-visuellen Materialsammlung.

Die Interaktion der Versuchspersonen kann zielführend oder zielabweichend sein und ihre Äußerungen können positiv oder negativ ausfallen sowie sich in ihrem Abstraktionsgrad unterscheiden. Pauschale Äußerungen beziehen auf das System insgesamt, konkrete Äußerungen auf spezifische Interface-Elemente.

Für die Detektion und Spezifikation von Usability-Problemen sind vor allem zielabweichende Interaktionen und konkrete, negative Äußerungen relevant, die sich auf spezifische Interface-Elemente beziehen, ggf. ergänzt durch pauschale Äußerungen zum Gesamtsystem. Alle anderen Interaktionen und Äußerungen sind für die Usability-Problematik irrelevant. Sie können sich aber in anderen Kontexten als hilfreich erweisen, z.B. im Rahmen qualitativer Studien zur User Experience (UX) oder zur Systemakzeptanz.

6 Datenanalyse und -interpretation

6.1 Zielsetzung und Vorgehensweise

Die audio-visuellen Daten der Materialsammlung sind die Rohdaten des Tests und bilden den Gegenstand der Datenanalyse. Das Ziel der Analyse besteht in der Detektion von Usability-Problemen und der Beantwortung der

Fragestellungen, die damit im Zusammenhang stehen (vgl. Kapitel 4.1). Des Weiteren sollen erste Ideen für die Optimierung des Systems und zur Beseitigung der Probleme erhoben werden.

Audio-visuelle Daten dieser Art werden durch Kodierer*innen ausgewertet, die das Material sichten und in Bezug auf mögliche Usability-Probleme interpretieren. Derartige Interpretationen müssen zwangsläufig auf Indikatoren basieren, da ein theoretisches Konstrukt wie „Usability-Problem“ nicht per se beobachtbar ist. Indikatoren zeigen sich im Verhalten der Testperson sowie in ihren Äußerungen während des Lauten Denkens bzw. während eines Interviews.

Die Reliabilität der Kodierungsergebnisse hängt maßgeblich von der Auswertungsobjektivität ab. Damit diese möglichst hoch ausfällt, sollte bei jeder Kodierung identisch vorgegangen werden. Dies lässt sich mit qualitativen Verfahren erreichen, die ein Prozessmodell für die Durchführung der Auswertung beinhalten.

Ein derartiges Prozessmodell stellt die **„Inhaltsanalyse mit induktiver Kategorienbildung“** für verbale Daten nach Mayring (2015) zur Verfügung. (vgl. Abbildung 8).

Die Analyse ist theoriegeleitet und erfolgt für einen festgelegten Untersuchungsgegenstand zur Erreichung eines vorgegebenen Ziels. Ihre zentrale Idee besteht darin, das vorhandene sprachliche Material sukzessive durchzuarbeiten und daraus induktiv Kategorien abzuleiten, die sich möglichst auf dem gleichen Abstraktionsniveau befinden. Dafür werden Äußerungen herausgefiltert, die vordefinierten Selektionskriterien genügen. Diese Aussagen werden fortlaufend den Kategorien zugewiesen, die ihrerseits kontinuierlich aus dem Material gewonnen werden. Redundante Beschreibungen, Ausschmückungen und nichtzielführende, irrelevante Äußerungen sowie synonyme Formulierungen, die das Material enthält, werden dabei eliminiert.

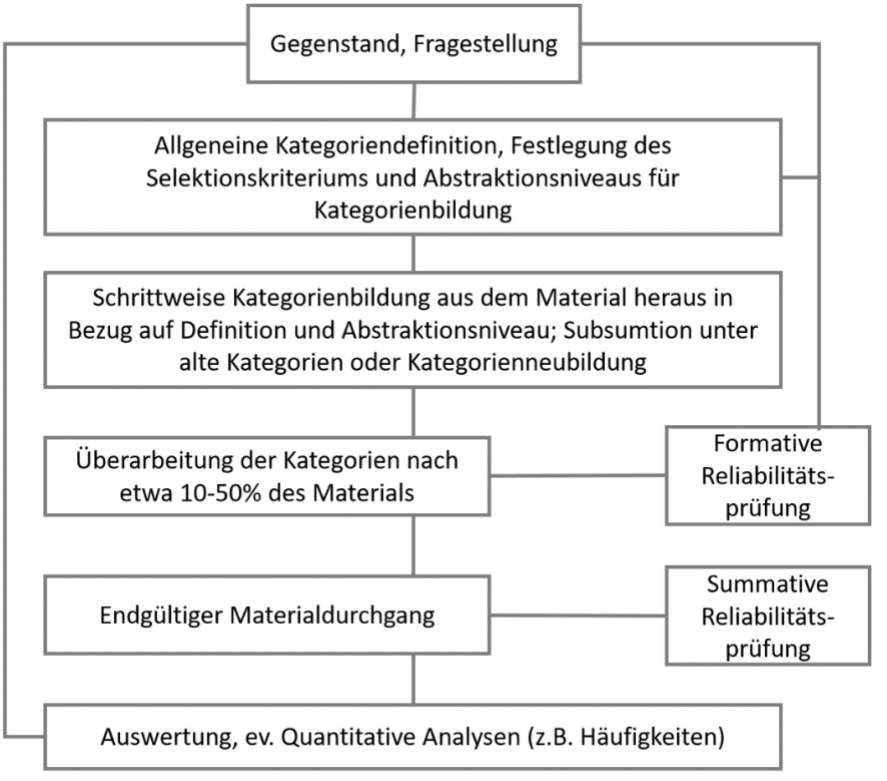

Abbildung 8: Vorgehensmodell der Inhaltsanalyse mit induktiver Kategorienbildung (Mayring, 2005, S. 472).

Der Verarbeitungsprozess wird durch zwei Arten der Reliabilitätsprüfung optimiert und abgesichert:

- **Formative Prüfung:** Nachdem 10% bis 50% des Materials gesichtet wurden, wird kontrolliert, ob die abgeleiteten Kategorien eine adäquate Struktur bilden oder überarbeitet werden müssen. Änderungen werden insbesondere notwendig, wenn erkannt wird, dass die Selektionskriterien nicht passend gewählt wurden oder dass sich nicht alle Kategorien auf dem gleichen Abstraktionsniveau befinden. Durch entsprechende Änderungen können Kategorien entfallen, hinzukommen, zusammengeführt oder umbenannt werden. Ist dies der Fall, wird das bereits analysierte Material neu zugewiesen und die veränderte Struktur für die Kodierung des verbleibenden Materials eingesetzt.

- **Summative Prüfung:** Zum Abschluss der Analyse wird das Material nochmals durchgegangen, um ggf. finale Änderungen vorzunehmen.

Das Ergebnis dieser systematischen Vorgehensweise ist ein strukturiertes, konsistentes und stark verdichtetes Inhaltsmodell des Materials. Dieses qualitative Modell lässt sich durch quantitative Daten ergänzen, indem pro Kategorie die Häufigkeit der Aussagen, die unter ihr subsumiert sind, ermittelt wird.

6.2 Inhaltsanalytische Kodierung des Lauten Denkens

Da Lautes Denken eine Fülle verbaler Daten liefert, ist die Inhaltsanalyse nach Mayring (2015) sehr gut für deren Strukturierung und Aggregierung geeignet. Allerdings bestehen für Usability-Tests einige Charakteristika, die zu berücksichtigen sind.

Eine Besonderheit betrifft zunächst die Art des zur Verfügung stehenden Materials. Bei einer herkömmlichen Inhaltsanalyse sind dies verbale Daten in gesprochener Sprache oder in Textform. Da bei einem Usability-Test nicht nur das laute Denken der Proband*innen aufgezeichnet wird, sondern auch ihre Interaktion mit dem System, steht für die Kodierung zusätzlich Bildmaterial zur Verfügung, das behaviorale Indikatoren für Usability-Probleme enthalten kann.

Eine weitere Besonderheit besteht in der Aufbereitung des Materials, die vor der eigentlichen Auswertung erfolgt. Bei einer herkömmlichen Inhaltsanalyse werden verbale Äußerungen, die in gesprochener Sprache vorliegen, zunächst verschriftet, damit die Kodierer*innen das Transkript flexibel und in selbstgewählter Geschwindigkeit durcharbeiten können. Bei der Kodierung der Aufzeichnung des Usability-Tests kann auf das Transkribieren verzichtet werden, da die Bild- und Tonaufzeichnung je nach Bedarf abgespielt und angehalten sowie vor- oder zurückgespult werden kann. Für die Detektion von Usability-Problemen können dabei nicht nur die Äußerungen der Versuchspersonen, sondern auch beobachtete Auffälligkeiten bei der Interaktion berücksichtigt werden.

Unter Berücksichtigung dieser Eigenheiten kann das von Mayring (2015) entwickelte Vergehen Schritt für Schritt auf die Datenanalyse von Usability-Tests übertragen werden (vgl. hierzu Abbildung 7).

Gegenstand der Analyse ist das getestete System (Software, Website, App, …), über das Material in Form der aufgezeichneten Videos vorliegt. Die **Fragestellungen**, die dem Usability-Test zugrunde liegen, sind in Kapitel 4.1 genannt und lassen sich zu zwei zentralen Fragen bündeln:

1. Welche Interface-Elemente sind bei der zielgerichteten Interaktion mit dem System von Usability-Problemen unterschiedlicher Schweregrade betroffen?
2. Welche Verbesserungshinweise ergeben sich aus Äußerungen oder zielabweichenden Verhaltensweisen der Proband*innen für ein Interface-Element?

Das Attribut „zielgerichtet“ in Frage 1 verweist darauf, dass die Interaktion zwischen Proband*innen und System bei der Bearbeitung von Prüfaufgaben untersucht wird. Die Frage nach „Verbesserungshinweisen“ impliziert, dass sich aus erfassten Äußerungen und beobachtetem Verhalten wohl eher selten endgültige Problemlösungen ableiten lassen, dass aber Anregungen in diese Richtung dokumentiert und für die Systemoptimierung geprüft werden sollten.

Die **allgemeine Kategoriendefinition** erfordert eine Entscheidung darüber, welche Arten von Kategorien bei der Analyse induktiv aus dem Material abgeleitet werden sollen. Die allgemeinen Kategorien, die bei einem Usability-Test eine Rolle spielt, sind bereits in beiden genannten Fragestellungen angesprochen; es geht um **Interface-Elemente**, für die Usability-Probleme bestehen und für deren Beseitigung ggf. Hinweise zu vorliegen. Von einem solchen Problem können alle Arten von Element betroffen sein - Icons, Symbole, Fenster, Menüs und ihre Unterpunkte, Benennungen (z.B. Funktionsbezeichnungen), Eingabefelder, Benachrichtigungen (z.B. Fehlermeldungen) usw. Die Aufgabe der Kodierer*innen besteht nun darin, diese allgemeinen Kategorien zu spezifizieren, also z.B. für die **allgemeine Kategorie** „Icon“ nach spezifischen Icons zu fahnden, deren Nutzung problematisch ist. So könnte es z.B. sein, dass das Diskettensymbol als Repräsentant für die Funktion „Datei speichern“ nicht allen Personen bekannt ist. In diesem Fall würde das Icon als **spezielle Kategorie** fungieren, unter der alle Probleme subsumiert werden, die mit ihm in Zusammenhang stehen.

Usability-Probleme äußern sich beim Lauten Denken in Form negativer Äußerungen oder als zielabweichendes Verhaltens.

Beispiele sind:

- **Aussagen zu Mängeln:** Schlechte Sichtbarkeit von Interface-Elementen (z.B. Verdeckungen), schlechte Lesbarkeit von Informationen (z.B. zu kleine Schrift, zu geringe Kontraste), Mehraufwände (Mehrfacheingaben), etc.
- **Fehlinterpretationen:** Missverständnis von Begriffen, Funktionen oder Icons, Fehlannahmen zum Systemzustand, falsche Erwartungen an Systemveränderungen nach Eingaben, etc.
- **Zielabweichendes Verhalten:** Abkommen vom kürzesten Lösungsweg einer Aufgabe, verlängerte oder gescheiterte Suchprozesse, Navigations- oder Orientierungsprobleme, etc.

Derartige Indikatoren fungieren als **Selektionskriterium**. Werden sie beim Kodieren entdeckt, so wird das betroffene Interface-Element als Kategorie klassifiziert und eine Problembeschreibung erstellt. Das **Abstraktionsniveau** derartiger Kategorien ist immer niedrig, da sie – anders als die allgemeine Kategorie „Interface-Element“ - eine **konkrete** Manifestation dieser Kategorie sind.

Korrespondierend zur Kategorie sollte auch die Beschreibung der Probleme, die mit ihr (also dem konkreten Interface-Element) verbunden sind, spezifisch und präzise sein. Sie sollte angeben:

- in welche spezifische Kategorie das Problem fällt (also bei welchem Interface-Element es auftritt),
- welcher Indikator in Form eines beobachteten Verhaltens und/oder einer Äußerung auf das Problem hindeutet,
- worin das Problem besteht,
- wie schwerwiegend das Problem ist,
- wie häufig das Problem auftritt und
- welche Optimierungsmöglichkeiten es für das Interface geben könnte, um das Problem zu beseitigen.

Durch die Erzeugung von Kategorien und Problembeschreibungen wird sukzessiv eine Sammlung von Usability-Problemen aufgebaut. Immer wenn die Kodierer*innen bei der Sichtung des Materials auf einen Indikator stoßen, wird

die Problemsammlung erweitert oder modifiziert. Dabei sind vier Fälle zu unterscheiden (vgl. Abbildung 9).

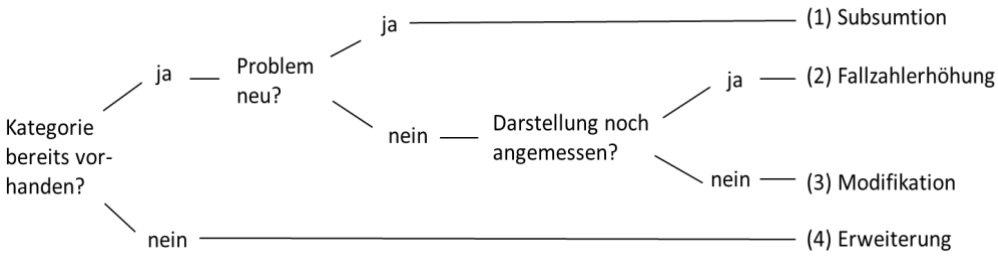

Abbildung 9: Vier Fälle der Erweiterung der Problemsammlung.

- **Fall (1):** Zeigt ein Indikator das Bestehen eines Problems an, wird zunächst per Abgleich mit der Problemsammlung geprüft, ob das betroffene Interface-Element schon als Kategorie berücksichtigt wurde. Trifft dies zu, wird im zweiten Schritt kontrolliert, ob der Indikator auf ein neues Problem für das Element hinweist. Falls ja, wird dafür eine Problembeschreibung erstellt und unter der bestehenden Kategorie **subsumiert**. Praktisch betrachtet bedeutet das, dass ein Kodierungsschema angelegt wird, das aus dem „alten“ Interface-Element und der Beschreibung des neuen Problems besteht. Dadurch erweitert sich also der Inhalt der Sammlung um einen Problemfall, ohne dass eine neue Kategorie entsteht.
- **Fall (2):** Weist der Indikator auf **kein** neues Problem für das Element hin, so liegt bereits eine Problembeschreibung vor und es wird geprüft, ob sie vor dem Hintergrund des neuen Falls noch angemessen ist. Falls ja, wird lediglich die **Fallzahl** für das Problem erhöht, ohne dass weitere Veränderungen vorgenommen werden.
- **Fall (3):** Ergibt sich bei der Prüfung ein Optimierungsbedarf, so ist sie zu **modifizieren**. Dafür kommen folgende Veränderungen in Betracht:
    - Umformulierung der Problembeschreibung, so dass sie auch den neuen Fall adäquat repräsentiert,
    - Ergänzung oder Ersatz der Beispiele für Äußerungen und/oder beobachteten Verhaltensweisen der Testpersonen,
    - Veränderung des eingeschätzten Schweregrades,
    - Ergänzung weiterer Optimierungsvorschläge.

Auch für den dritten Fall ist die Auftretenshäufigkeit des Problems zu erhöhen.

- **Falls 4:** Ist das interface-Element **noch nicht** als problematisch kodiert worden, muss die Problemdokumentation **erweitert** werden. Das betroffene Interface-Element wird zur neuen Kategorie deklariert und gemeinsam mit der Problembeschreibung, die auf Basis des Indikators zu erstellen ist, in die Problemsammlung integriert.

Auf die beschriebene Weise wird fortgefahren, bis die gesamte Materialsammlung durchgearbeitet wurde, also bis alle Videosequenzen, die die Bearbeitung der Prüfaufgaben durch die Versuchspersonen zeigen, kodiert und die dabei ermittelten Probleme in die Problemdokumentation aufgenommen wurden. Hierbei ergibt sich ein Unterschied zur Inhaltsanalyse nach Mayring (2015), bei der eine **formative Reliabilitätsprüfung** und Überarbeitung der Kategorien erfolgt, nachdem 10% bis 50% des Materials kodiert wurden. Dies ist bei der hier entwickelten Systematik nicht nötig, da bei jedem Auftreten von Indikatoren ein Abgleich mit den Einträgen der Problemsammlung vorgenommen wird, so dass etwaige Änderungen sich sofort umsetzen lassen. Ebenso wie bei Mayring sollte nach beendeter Sichtung des Materials allerdings final eine **summative Reliabilitätsprüfung** erfolgen, indem nochmals alle Einträge der Sammlung durchgegangen und nötigenfalls modifiziert werden. Ist es im Zuge der Kodierung zu Modifikationen gekommen, sollte dabei ein besonderes Augenmerk auf etwaige Veränderungen von Problembeschreibungen (Fall 3) gerichtet werden.

Zum Abschluss der Analyse können die erfassten Fälle **quantifiziert** werden, d.h. es kann ausgezählt werden, wieviele Probleme über alle Versuchspersonen und Prüfaufgaben hinweg entdeckt wurden und welche Interface-Elemente besonders von Mängeln betroffen sind.

6.3 Kodierungsschema und Schweregradskala

Um eine gute Reliabilität der Kodierungsergebnisse zu gewährleisten, sollte die Auswertungsobjektivität bei der Analyse der erhobenen Daten möglichst hoch sein. Dies kann durch eine Standardisierung der Vorgehensweise erreicht werden.

Einen wichtigen Beitrag dazu leistet das bereits vorgestellte inhaltsanalytische Prozessmodell, wenn es konsequent über alle Versuchspersonen und Prüfaufgaben umgesetzt wird. Eine weitere Erhöhung der Standardisierung kann dadurch erreicht werden, dass die zentralen Charakteristika von Usability-Problemen (vgl. Kapitel 6.2) in ein **Kodierungsschema** überführt werden, das die Kodierer*innen bei der Datenanalyse benutzen, um ein Problem zu erfassen und zu spezifizieren (vgl. Tabelle 5).

Tabelle 5: Kodierungsschema zur Erfassung von Usability-Problemen.

(01) Versuchsperson	Eintrag der Versuchspersonenkennung
(02) Kategorie	Vom Problem betroffenes Interface-Element
(03) Problem	Benennung und präzise Beschreibung des Problems
(04) Äußerung	Beispiel für eine Probandenäußerung zum Problem
(05) Beobachtung	Beispiel für beobachtetes, problematisches Verhalten
(06) Schweregrad	Ratingwert zur Einschätzung der Problemschwere
(07) Prüfaufgaben	Kennung der Prüfaufgaben
(08) Fallzahl	Auftretenshäufigkeit des Problems
(09) Quelle	Lautes Denken oder Interview
(10) Optimierungsideen	Probandenvorschläge zur Problembeseitigung
(11) Anmerkungen	Notizen zur Kodierung

Die Zeilen des Schemas dienen folgenden Eingaben:

- In **Zeile 1** wird die **Kennung der Versuchsperson** vermerkt, deren Video kodiert wird.
- **Zeile 2** ist für die Festlegung der **Kategorie** vorgesehen. Hier wird das **Interface-Element** angegeben, für das das Usability-Problem besteht. Das Element sollte durch eine adäquate **Benennung,** möglichst in Kombination mit einem **Screenshot**, referenziert werden.
- **Zeile 3** sollte eine treffende **Problembenennung** und eine möglichst knappe und präzise **Problembeschreibung** enthalten.
- Die **Zeilen 4 und 5** geben problembezogene, **wörtliche Zitate** von Äußerungen der Versuchsperson und **Beobachtungen der Kodierer*innen** bei der Sichtung des Materials wieder.
- **Zeile 6** enthält eine **kriterienbasierte Einschätzung** der Problemschwere durch den/die Kodierer*in.
- In **Zeile 7** wird die Kennung der Prüfaufgaben angegeben, bei denen die Versuchsperson das Problem hatte.

- In **Zeile 8** wird erfasst, wie **häufig** das Problem bei der Versuchsperson über alle Prüfaufgaben hinweg aufgetreten ist.
- **Zeile 9** nennt die **Quelle**, auf der die Kodierung beruht (Aufzeichnung des Lauten Denkens oder Interview).
- Hat die Testperson **Ideen zur Problembeseitigung** geäußert, so werden sie in **Zeile 10** angegeben.
- **Zeile 11** können die Kodierer*innen frei für etwaige **Notizen** und **Kommentare** verwenden, z.B. auch für Ideen zur Systemoptimierung, die von ihnen selber und nicht von den Versuchspersonen stammen

Das Schema wird für die Problemerfassung und -beschreibung über alle Versuchspersonen und Prüfaufgaben hinweg angewendet und jedes ausgefüllte Schema wird gemäß der vier Fälle in Abbildung 9 behandelt, um die Problemsammlung sukzessive aufzubauen und ggf. zu modifizieren.

Zur Bestimmung des **Schweregrades** eines Problems kann eine **Ratingskala** verwendet werden. Derartige Skalen sind meist 3- bis 7-stufig und sollten Kriterien zur Vergabe des Urteils spezifizieren. Abbildung 10 zeigt ein Beispiel für eine 3-stufige Skala sowie die zugehörigen Kriterien für die Beurteilung der Schwere des Problems.

Gemeinsam mit dem inhaltsanalytischem Prozessmodell ermöglichen das **Kodierungsschema** und die **Schweregradskala** eine weitgehend standardisierte Datenanalyse und unterstützen eine **systematische**, **nachvollziehbare** Auswertung und Interpretation der Materialsammlung durch die Kodierer*innen.

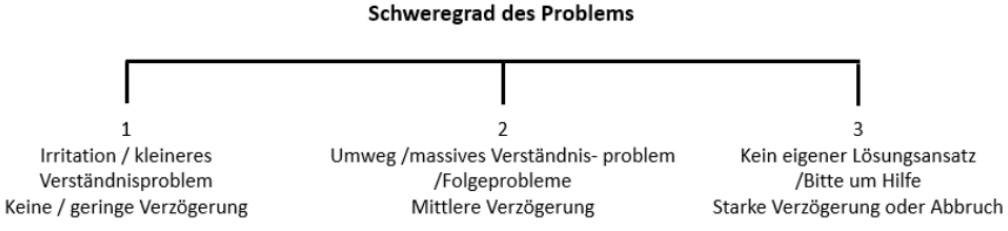

Abbildung 10: Beispiel einer 3-stufigen Ratingskala für die Bestimmung des Schweregrads von Usability-Problemen.

6.4 Mehrfachkodierung und Reliabilität

Für die Datenanalyse von Usability-Tests ist die Erstellung von zwei bis maximal drei Kodierungen pro Datensatz empfehlenswert. Zwei Gründe sprechen für eine Mehrfachkodierung. Zum einen steigt dadurch die Wahrscheinlichkeit, bestehende Usability-Probleme zu identifizieren, zum anderen lässt sich die Reliabilität der Interpretation von Äußerungen und Verhaltensweisen als Indikatoren nur dann überprüfen, wenn mindestens zwei Kodierungen vorliegen. Von mehr als drei Kodierungen ist allerdings aufgrund des damit verbundenen Auswertungsaufwands in der Regel abzuraten.

Das effektivste und effizienteste Ergebnis lässt sich durch die Umsetzung des **„Vier-Augen-Prinzips“** bei der Kodierung erzielen, wobei man sich in der Praxis allerdings häufig mit nur einer*einem Kodierer*in zufriedengibt. Ist dies der Fall, sollte die Kodierung durch diese Person nach einem längeren Zeitraum (idealerweise mindestens 1 Woche) in veränderter Reihenfolge der Proband*innen wiederholt werden, um Gedächtniseffekte zu reduzieren.

Bei einer Mehrfachkodierung entstehen zwei Problemsammlungen, die sich so gut wie immer in Umfang und Inhalt unterscheiden (vgl. Tabelle 6).

Die Beispieltabelle zeigt die beiden Interface-Elemente A und B, mit den Problemen X und Y, den Werten 0 oder 1 für deren Detektion sowie die Schweregrade (1 bis 3). Das Problem X ist in Kodierung 1, nicht aber in Kodierung 2 diagnostiziert worden. Im ersten Fall wurden also Aussagen mindestens einer Testperson als Indikator für Problem X interpretiert, im zweiten Fall nicht. Problem Y hingegen wird in beiden Kodierungen diagnostiziert, allerdings ohne, dass Einigkeit über dessen Schweregrad besteht.

Das Ergebnis der Kodierung ist umso verlässlicher (reliabler), je größer die Übereinstimmung zwischen den beiden Problemsammlungen hinsichtlich der identifizierten Usability-Probleme und der zugeordneten Schweregrade ist.

Tabelle 6: Gemeinsamkeiten und Unterschiede zweier Kodierungen.

Interface-Element	A		
Problem	X		
		Kodierer 1	**Kodierer 2**
Detektion		1	0
Schweregrad		3	
Interface-Element	B		
Problem	Y		
		Kodierer 1	**Kodierer 2**
Detektion		1	1
Schweregrad		3	2
Interface-Element …	…..	…..	….

Die einfachste Form der Übereinstimmungsbetrachtung ist die Berechnung der prozentualen Übereinstimmung:

*$Prozent_{identisch}$ = Anzahl übereinstimmender Problembeurteilungen/Gesamtzahl der Beurteilungen * 100%*

Enthält die Materialsammlung z.B. 250 Aussagen, die in zwei Kodierungen durchgearbeitet werden, so beträgt die Gesamtzahl der Urteile 500. Werden in beiden Kodierungen übereinstimmend 50 der Aussagen als Indikatoren für ein Usability-Problem interpretiert und 425 übereinstimmend als Indikatoren abgelehnt, so ergibt sich als prozentuale Übereinstimmung:

$$Prozent_{identisch} = (50 + 425) / 500 * 100\% = 95\%$$

Diese einfache Berechnung beachten allerdings ***nicht explizit***,

- dass einige Urteile **nicht** übereinstimmen (im Beispiel 25) sowie
- dass einige der Übereinstimmungen **zufällig** entstanden sein können.

Beides wird durch ein statistisches Maß der Reliabilität berücksichtigt, durch den sog. **Kappa-Koeffizienten (K)** nach Cohen (1960). Anhang 2 zeigt, wie Kappa ermittelt wird, erläutert den Rechenweg an einem Beispiel und stellt einen Link zu einem Online-Rechner für die Kalkulation von Kappa zur Verfügung.

Für die Interpretation des Koeffizienten hat sich mittlerweile eine Konvention nach Landis und Koch (1977) etabliert. Danach wird ein Kappa **größer .80** als

sehr gut oder fast perfekt eingestuft, ein Wert **größer .60** als gut oder substantiell (vgl. Tabelle 7).

Tabelle 7: Interpretation von Cohens Kappa nach Landis und Koch (1977).

Values	Interpretation
Smaller than 0.00	Poor Agreement
0.00 to 0.20	Slight Agreement
0.21 to 0.40	Fair Agreement
0.41 to 0.60	Moderate Agreement
0.61 to 0.80	Substantial Agreement
0.81 to 1.00	Almost Perfect Agreement

Wird Kappa für die Bestimmung der Reliabilität einer Usability-Analyse berechnet, so sollte ein Wert größer 0.60 erreicht werden. Dieser Wert kann als Akzeptanzschwelle angesehen werden, die zu überschreiten ist, damit die Kodierung nicht wiederholt werden muss. Je nach Zielsetzung der Analyse kann jedoch auch ein anderer Schwellenwert gewählt werden, der liberaler oder konservativer ist.

Verfehlt Kappa die gewählte Akzeptanzschwelle, so bestehen Zweifel daran, dass die Messgenauigkeit und die Verlässlichkeit der Ergebnisse den gesetzten Ansprüchen des Tests genügen.

In diesem Fall gibt es **zwei Möglichkeiten**:

1. Die Ergebnisse werden verworfen und der Test wird ggf. wiederholt. Diese Option ist aus Kostengründen wenig attraktiv.
2. Wurde die Auswertung von zwei Kodierer*innen unabhängig voneinander vorgenommen, so kann eine „Kodierkonferenz“ (Mayring, 2015, S. 125) durchgeführt werden, in der die Problemdokumentationen verglichen werden, um die Anzahl divergierender Urteile zu reduzieren und gemeinsam eine konsolidierte Dokumentation zu erstellen.

Auch bei Einhaltung der Akzeptanzschwelle ist eine solche Kodierkonferenz zur Konsolidierung der Ergebnisse sinnvoll, da sie hilft, Missverständnisse auszuräumen und Kodierungsfehler zu reduzieren. Als Vorbereitung auf die Konferenz sollten beide Kodierer*innen ihre Problemsammlung auf identische

Weise sortieren, z.B. nach den Pseudonymen der Versuchspersonen in alphabetischer Folge. Da die Prüfaufgaben randomisiert dargeboten wurden, ist die Reihenfolge innerhalb der Problemsammlungen unterschiedlich, so dass auch diese sortiert werden sollten, am besten in der Reihenfolge der spezifischen Kategorien, so dass die problembetroffenen Interface-Element nacheinander diskutiert werden können. Um eine effiziente Sortierung zu ermöglichen, empfiehlt es sich das Kodierungsschema als Exceldatei anzulegen.

Für die Ermittlung der Reliabilität ist zu berücksichtigen, dass dafür ***alle*** Äußerungen berücksichtigt werden müssen, also nicht nur jene, bei denen ein Problem diagnostiziert wurde, sondern auch jene, bei denen die Kodierer*innen übereinstimmend zu dem Schluss gekommen sind, dass die Äußerungen kein Usability-Problem betreffen. Der dadurch entstehende Aufwand ist viel höher als der, der für die Problemdetektion nötig ist. Aus diesem Grund wird in der Praxis fast immer auf eine statistische Ermittlung der Reliabilität gänzlich verzichtet oder es wird lediglich die Übereinstimmung bei der Diagnose der Probleme bestimmt; analog zu der oben beschriebenen prozentualen Übereinstimmung, also gemäß:

$$Prozent^*_{identisch} = Anzahl\ übereinstimmender\ Problembeurteilungen\ /\ Gesamtzahl\ der\ Problembeurteilungen\ *\ 100\%$$

$Prozent^*_{identisch}$ bezieht also lediglich die Anzahl der insgesamt entdeckten Probleme ein, nicht die Gesamtzahl der Beurteilungen. Werden z.B. insgesamt 50 Probleme identifiziert, von denen 40 Fälle gemeinsam in beiden Kodierungen enthalten sind, ist $Prozent^*_{identisch} = 40 / 50 * 100\% = 80\%$.

Da dieser Wert nicht alle Aussagen berücksichtigt, kann er nur als grobe Abschätzung gelten und sollte entsprechend vorsichtig interpretiert werden. Außerdem ist davon auszugehen, dass die so ermittelte Übereinstimmung zu hoch ausfällt, weil sie nicht um den erwähnten Zufallseinfluss korrigiert wird.

Weil die so ermittelte Übereinstimmung nicht um den erwähnten Zufallseinfluss korrigiert wird, ist davon auszugehen, dass sie meist zu hoch ausfällt. Außerdem kann sie nur als grobe Schätzung fungieren und sollte entsprechend vorsichtig interpretiert werden. Ist es aus Mangel an Ressourcen oder Zeit allerdings nicht machbar, die Reliabilität nach Kappa zu bestimmen,

gibt $Prozent*_{identisch}$ zumindest eine grobe Orientierung hinsichtlich der Verlässlichkeit der Problemdetektion.

7 Problemdokumentation, Projektbericht und Optimierungsvorschläge

Die Ergebnisse sind final in einer umfassenden Problemdokumentation festzuhalten, die die erstellten Kodierungsschemata zusammenfasst. Ist eine Kodierkonferenz durchgeführt worden, so enthält die Dokumentation die konsolidierten Ergebnisse beider Kodierer*innen. Auf ihrer Basis kann abschließend ein Projektbericht erstellt werden, der nicht nur die Testergebnisse umfassen sollte, sondern auch die Fragestellung und Methodik der durchgeführten Evaluation (vgl. Anhang 3 für eine beispielhafte Gliederung).

Für den nachfolgenden Zyklus des UCD-Prozesses ist es von zentraler Bedeutung, zu entscheiden, welche der gefundenen Probleme vorrangig zu beheben sind. Dafür können sie in Abhängigkeit vom beurteilten Schweregrad in eine Rangordnung gebracht werden. Schwere Probleme sind zuerst zu beseitigen, gefolgt von mittelschweren und letztlich auch den leichteren, wenn Zeit und Ressourcen dies zulassen. Für die die Optimierung des Systems sollten außerdem die Vorschläge in Betracht gezogen und diskutiert werden, die in den Kodierungsschemata enthalten sind.

8 Objektivität, Reliabilität und Validität Lauten Denkens

Die Übertragbarkeit der „klassischen“ psychometrischen Gütekriterien – Objektivität, Reliabilität und Validität – auf qualitative Methoden, wie die des Lauten Denkens, wird von vielen Forscher*innen bezweifelt. Einen Lösungsansatz schlagen Sedlmeier und Renkewitz (2008) mit dem Konzept der sog. „prozeduralen Reliabilität“ vor. Sie hat „den Zweck, die Vertrauenswürdigkeit von Daten und Interpretationen zu erhöhen, indem man die Prozedur weitgehend standardisiert.“ (S. 763). In dem hier vorgelegten Kompendium wurde deshalb eine **Vorgehensweise** entwickelt, die darauf abzielt, für alle drei Objektivitätsaspekte (Testdurchführung sowie Auswertung und Interpretation der erhobenen Daten) einen hohen Grad an **Standardisierung** für formative Usability-Tests zu erreichen.

Zur Gewährleistung der **Durchführungsobjektivität** wurde ein Leitfaden zur Datenerhebung erstellt (vgl. Tabelle 4) und um ausführliche Anleitungen zur

Gestaltung simultanen und retrospektiven Lauten Denkens ergänzt (vgl. Abschnitt 4.3.1 und 4.3.2). Zur Standardisierung der Kommunikation zwischen Versuchsleitung und Testpersonen wurden **Textbausteine** für Begrüßung, Instruktion und demografische Fragen entworfen und um den Text für eine Einverständniserklärung zur Datenaufzeichnung und -verwendung ergänzt. Des Weiteren wurde ein Verfahren zur Pseudonymisierung der Versuchspersonen vorgestellt (zu allem vgl. Anhang 1). Werden diese Materialien zusammen mit dem Leitfaden bei der Versuchsdurchführung konsequent eingesetzt, sollte sich die Datenerhebung in jedem Test analog gestalten und von den durchführenden Personen weitgehend unabhängig sein.

Um eine möglichst hohe **Auswertungsobjektivität** herzustellen, wurde das Prozessmodell der Inhaltsanalyse mit induktiver Kategorienbildung (vgl. Abbildung 8) nach Mayring (2015) auf die Analyse verbaler Daten, die beim Lauten Denken anfallen, übertragen und hierfür angepasst. Im Zentrum der Anpassung steht die Idee, Interface-Elemente, bei denen sprachliche oder behaviorale Indikatoren auf Probleme hindeuten, als Kategorien zu verwenden und detektierte Probleme darunter zu subsumieren. Ergänzt wird dieses regelhafte Vorgehen um ein **Kodierungsschema**, das auf einer etablierten Definition des Konstrukts „Usability-Problem“ (Sarodnick & Brau, 2011) basiert und eine standardisierte Erfassung und Beschreibung derartiger Probleme unterstützt. Außerdem wurde ein 3-stufiges **Ratingverfahren** vorgeschlagen, das die Einschätzung des Schweregrades auf Basis vordefinierter Kriterien vereinheitlicht. Inhaltsanalytisches Vorgehensmodell, Kategorisierungsschema und Ratingverfahren resultieren in einer strukturierten Dokumentation von Problemen, Schweregraden und Optimierungsvorschlägen mit direktem Bezug zum Interface des evaluierten Systems.

Um möglichst viele Usability-Probleme zu entdecken und gleichzeitig eine Einschätzung der Reliabilität der Ergebnisse zu ermöglichen, wird empfohlen, nach dem **Vier-Augen-Prinzip** zu verfahren und zwei Kodierer*innen bei der Analyse einzusetzen. Dadurch entstehen zwei Problemdokumentationen, die in der Regel durch Unterschiede aufgrund verschiedener Interpretationen der Proband*innenaussagen voneinander abweichen. Zur Verbesserung der **Interpretationsobjektivität** sollte einem Vorschlag von Mayring (2015) gefolgt

werden, der dazu rät, eine **Kodierkonferenz** durchzuführen, um Divergenzen in den Kodierungen zu diskutieren und konsensual zu aufzuheben.

Inwieweit die durch Standardisierung erzeugte Objektivität hinsichtlich Testdurchführung, Datenauswertung und Interpretation zu einer angemessenen **Reliabilität** der Ergebnisse beiträgt, kann durch die Berechnung des Kappa-Koeffizienten nach Cohen (1960) und seiner Bewertung nach Landis und Koch (1977) beurteilt werden. Liegt sein Wert über einer festgelegten Akzeptanzschwelle, wie z.B. 0.60, so gilt die Übereinstimmung als substantiell und kann als ausreichend für die Interpretation der Daten angesehen werden.

Während sich eine hohe Objektivität der Methodik des Lauten Denkens durch Standardisierung weitgehend erreichen lässt, und ihre Reliabilität statistisch berechnet werden kann, gestaltet sich die Beurteilung ihrer Validität als schwierig. Die üblichen Validitätskriterien für quantitative Daten, wie z.B. die Bezugnahme auf ein Außenkriterium oder die Untersuchung der Vorhersage- oder der Konstruktvalidität, lassen sich nur schwer auf qualitative Daten übertragen und statistisch absichert quantifizieren. Zwar ist es vorstellbar, eine Validierung an Außenkriterien vorzunehmen und die Vorhersagevalidität zu prüfen, indem zusätzlich zu den verbalen Daten quantitative Daten[16] erhoben und Korrelationen berechnet werden, doch findet dies in der Praxis keine Anwendung.

Trotz dieser Einschränkungen ist es auch für die praktische Anwendung wichtig, die Validität der Ergebnisse Lauten Denkens zu prüfen und möglichst zu erhöhen. Hierfür schlagen Sedlmeier und Renkewitz (2008) Folgendes vor:

- Zum einen können die Ergebnisse des Lauten Denkens **„kommunikativ validiert“** werden, indem man sie den Testpersonen vorlegt und diese dazu Stellung nehmen. Im hier vorgestellten Ansatz könnte dies zumindest in Teilen durch Interviews geleistet werden, in denen die Testpersonen bestätigen oder dementieren, dass bestimmte Äußerungen oder Verhaltensweisen, die von der Versuchsleitung protokolliert wurden, auf Usability-Probleme hinweisen. Prinzipiell ist

[16] Derartige Daten könnten z.B. Fehlerhäufigkeiten, die Länge von Navigationswegen oder Bearbeitungszeiten sein. Zu erwarten wären hierfür positive Korrelationen mit der Anzahl der Usability-Probleme, die die Testpersonen entdecken.

es natürlich auch möglich, auf diese Weise die gesamten Ergebnisse des Tests validieren zu lassen, was aber aufgrund des recht hohen Aufwands in der Praxis bislang nur selten Anwendung findet. Außerdem wäre zu bedenkend, dass Antworttendenzen in Richtung sozialer Erwünschtheit bestehen könnten, die die Ergebnisse verzerren würden.

- Zum anderen kann das aus der Landvermessung bekannte Verfahren der **Triangulation** übernommen und zur Validierung der diagnostizierten Usability-Probleme herangezogen werden. Triangulation bedeutet, einen Forschungsgegenstand aus mindestens zwei verschiedenen Perspektiven zu betrachten. Von den vier Triangulationsarten ist vor allem die sog. „Forschertriangulation" im Rahmen des Lauten Denkens anwendbar. Hierbei werden verschiedene Beobachter eingesetzt, um Verzerrungen bei der Auswertung und Interpretation der Daten aufzudecken. Wird bei der inhaltsanalytischen Kodierung das Vier-Augen-Prinzip praktiziert, ist eine solche Triangulation gegeben und sollte zu einer Erhöhung der Validität beitragen.

Zusammenfassend lässt sich für die Methode des Lauten Denkens festhalten, dass die Formen der Standardisierung, die in diesem Kompendium vorgestellt wurden, nicht nur geeignet sind, die Objektivität und Reliabilität formativer Usability-Tests zu erhöhen, sondern auch zu einer Steigerung der Validität beitragen können – selbst wenn dies nicht im selben Maße überprüfbar ist wie für Methoden, deren Ergebnisse quantitativer Natur sind.

9 Literaturverzeichnis

Bevan, N., Kirakowski, J., & Maissel, J. (1991). What is usability? In H. J. Bullinger (Ed.), Human Aspects in Computing. Design and Use of Interactive Systems and Work with Terminals. *Proceedings of the 4th International Conference on Human-Computer Interaction* (pp. 651-655). Stuttgart, Germany: Elsevier.

Brooke, J. (1996). SUS: A "quick and dirty" usability scale. In P. W. Jordan, B. Thomas, B. Weerdmeester & I. L. McClelland (Eds.), *Usability Evaluation in Industry* (pp. 189-194). London, UK: Taylor & Francis.

Chapanis, A. (1981). Evaluating Ease of Use. *Unpublished Manuscript prepared for IBM,* Boca Raton, Fl., USA.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1), 37-46.

Ergonomie der Mensch-System-Interaktion - Teil 11: *Gebrauchstauglichkeit: Begriffe und Konzepte* (ISO 9241-11:2018); Deutsche Fassung EN ISO 9241-11:2018. Berlin: Beuth Verlag.

Ergonomie der Mensch-System-Interaktion - Teil 210: *Menschzentrierte Gestaltung interaktiver Systeme* (ISO 9241-210:2019); Deutsche Fassung EN ISO 9241-210:2019. Berlin: Beuth Verlag.

Ericsson, K.A. & Simon, H.A. (1980). Verbal Reports as Data. *Psychological Review*, 87(3), 215-251.

Faulkner, L. (2003). Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, & Computers*, 35(3), 379-383.

Hornbaek, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *Human Computer Studies*, 64, 79-102.

Kirakowski, J. (1996). The Software Usability Measurement Inventory: Background and Usage. In P. W. Jordan, B. Thomas, I. L. McClelland, & B. Weerdmeester, *Usability Evaluation in Industry* (pp. 189-194). London, UK: Taylor & Francis.

Kirakowski, J., & Corbett, M. (1993). SUMI: the Software Usability Measurement Inventory. *British Journal of Educational Technology*, *24*(3), 210–212.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174. https://doi.org/10.2307/2529310

Lewis, J. R. (2014). Usability: Lessons Learned … and yet to be Learned. *International Journal of Human-Computer Interaction*, 30(9), 663-684.

Mayring, P. (2015). *Qualitative Inhaltsanalyse. Grundlagen und Techniken.* Weinheim und Basel: Beltz.

Nielsen, J., & Landauer, T. K. (1993, May). A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems* (pp. 206-213).

Pataki, K., Sachse, K., Prümper, J., & Thüring, M. (2006). ISONORM 9241/10-S: Kurzfragebogen zur Software-Evaluation. In *Berichte über den 45. Kongress der Deutschen Gesellschaft für Psychologie* (pp. 258-259).

Perfetti, C. (2001). *Eight is not enough*. Retrieved from https://articles.uie.com/eight_is_not_enough/

Prümper, J. (1997). Der Benutzungsfragebogen ISONORM 9241/10: Ergebnisse zur Reliabilität und Validität. In R. Liskowski, B.M. Velichkovsky & W. Wünschmann (Hrsg.), *Software-Ergonomie'97 - Usability Engineering: Integration von Mensch-Computer-Interaktion und Software-Entwicklung* (pp. 253-262). Stuttgart: Teubner.

Sarodnick, F. & Brau, H. (2011). *Methoden der Usability Evaluation. Wissenschaftliche Grundlagen und praktische Anwendung*. Hogrefe: Bern.

Sedlmeier, P. & Renkewitz, F. (2008). *Forschungsmethoden und Statistik in der Psychologie*. Pearson Studium: München.

Streitz, N. A. (1986, April). Cognitive ergonomics: An approach for the design of user-oriented interactive systems. In *Seminar of The International Union of Psychological Science (IUPsyS) on Man-computer interaction research (MACINTER-I): Proceedings of the first network* (pp. 21-33).

Tractinsky, N. (2017). The Usability Construct: A Dead End? *Human-Computer Interaction*, 00, 1-47.

Tullis, T. S. (2019). Nigel Bevan: An overview of his contributions to usability and UX. *Journal of Usability Studies*, 14(3), 134-144.

Virzi, R. A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors*, 34, 457-468.

Anhang 1: Materialien für die Testdurchführung

Im Folgenden werden aufeinander abgestimmte Vorschläge zur Gestaltung von Materialien in der Reihenfolge vorgestellt, die für die Testdurchführung benötigt wird. In eckige Klammern eingeschlossene Textstellen können verwendet werden, um das jeweilige Material so anzupassen, wie in den Erläuterungen beschrieben wird.

Begrüßung und Einleitung

Vielen Dank für Ihre Teilnahme an unserer Untersuchung. Das Ziel ist es, Probleme aufzudecken, die bei der Nutzung [_*B1*_] auftreten. In der Untersuchung werden Ihnen Aufgaben mit der Bitte vorgelegt, sie zu bearbeiten. Von der Bearbeitung wird für die Auswertung eine Bild- und Tonaufzeichnung erstellt.

Für die Erhebung, Nutzung und Speicherung der Daten, die in der Untersuchung erhoben werden, benötigen wir Ihr Einverständnis.

Bitte lesen Sie die vorliegende Einverständniserklärung gründlich durch und unterschreiben Sie sie, wenn Sie keine Einwände haben. Fragen hierzu beantworten wir Ihnen gerne.

Erläuterung:

[_*B1*_]: Name und Art des Systems einfügen und ggf. Kurzvorstellung des Systems.

Aufklärung und Einverständniserklärung

Ich bin damit einverstanden, dass die im Rahmen dieser Untersuchung erhobenen Daten ohne Nennung meines Namens (pseudonymisiert) gespeichert und ausgewertet werden. Personenbezogene Daten, wie z.B. diese Einverständniserklärung, werden streng vertraulich behandelt und getrennt von den erhobenen Daten aufbewahrt. Ihre Weitergabe an Dritte ist ausgeschlossen.

Meine Teilnahme an der Untersuchung erfolgt freiwillig. Daher kann ich meine Teilnahme zu jedem Zeitpunkt ohne Angaben von Gründen abbrechen, ohne dass mir daraus Nachteile entstehen. Bei Widerruf des Einverständnisses werden alle erhobenen Daten gelöscht.

Mir ist bewusst, dass die Auswertung in Form von Berichten, Präsentationen oder Publikationen dokumentiert wird, allerdings ohne, dass Rückschlüsse auf meine Person möglich sind.

Alle pseudonymisierten Daten werden für die Dauer von [_*E1*_] archiviert, ohne dass sie Dritten zugänglich sind. Nach Ablauf dieser Zeit werden die Daten gelöscht. Wenn Sie vor Ablauf der Frist die Löschung Ihrer Daten wünschen, wenden Sie sich bitte an [_*E2*_].

Während der heutigen Datenerhebung erfolgt zu Auswertungszwecken eine Bild- und Tonaufzeichnung.

Ich bestätige durch meine Unterschrift, dass ich die Aufklärung verstanden habe und mich mit der Nutzung und Veröffentlichung meiner pseudonymisierten Daten sowie der Aufzeichnung der Datenerhebung einverstanden erkläre.

Vorname, Name: Ort:
Datum:

Unterschrift

Erläuterung:

[_*E1*_]: Zeitraum angeben.

[_*E2*_]: Kontaktperson mit Namen, Email und ggf. Telefonnummer angeben.

Pseudonymisierung

Um die erhobenen Daten nicht unter Ihrem Namen, sondern anonymisiert speichern zu können, muss ein Pseudonym erzeugt werden. Sie können uns ein eigenes Pseudonym nennen, aus Datenschutzgründen empfehlen wir aber ein synthetisches Verfahren, das eine abstrakte Kennung erzeugt. Wenn Sie damit einverstanden sind, nennen Sie nun bitte:

1. den dritten Buchstaben Ihres Vornamens (z.B. „N“ für Günther)
2. den letzten Buchstaben Ihres Nachnamens (z.B. „O“ für Schmidt)
3. den ersten Buchstaben des Vornamens Ihrer Muttern (z.B. „I“ für Irene)
4. den ersten Buchstaben Ihres Geburtsortes (z.B. „F“ für Freiburg)

5. den Tag Ihres Geburtsdatums (z.B. „13“ für 13.07.1986 oder „08“ für 08.07.1986; den Tag also bitte zweistellig angeben)
6. Ihre Position in der Geschwisterreihenfolge (z.B. „2“, wenn Sie als zweite*r von drei Geschwistern geboren wurden, „1“ oder „2“ bei Zwillingen, je nachdem wer früher zur Welt kam und „0“ falls Sie Einzelkind sind).

(Im Beispiel ergibt sich „NTIF130“ als Pseudonym für ein Einzelkind.)

Wenn Sie die Löschung Ihrer Daten vor Ablauf der Archivierung veranlassen möchten, geben Sie dafür bitte das Pseudonym an, unter dem die Daten archiviert wurden.

Demografische Angaben

Für die Datenauswertung benötigen wir noch einige Angaben zu Ihrer Person:

1. Welches Geschlecht haben Sie? (Weiblich, männlich, divers / keine Angabe)
2. Wann sind sie geboren? (Geburtsmonat und -jahr / keine Angabe)
3. Was ist Ihre Muttersprache?
4. Welchen höchsten allgemeinbildenden Schulabschluss haben Sie?
5. Welche beruflichen Ausbildungsabschlüsse haben Sie?
6. Was ist Ihre aktuelle Erwerbssituation?

Erläuterung: Die ersten beiden Fragen bilden das Minimum demografischer Angaben, die anderen vier sind in den meisten demografischen Fragebögen enthalten. Weitere Fragen können bei Bedarf ergänzt werden, wobei nur solche Ergänzungen vorgenommen werden sollten, die für die Analyse und Interpretation der Daten im Vorfeld als relevant erscheinen, insbesondere wenn es darauf ankommt, Personen zu testen, die einem vorgegebenen Nutzerprofil oder einer Persona entsprechen.

Die demografischen Angaben sollten unter dem erzeugten Pseudonym abgespeichert werden.

Instruktion zum Lauten Denken

Mit dieser Untersuchung sollen Probleme ermittelt werden, die Personen haben, die das System benutzen. Gegenstand der Untersuchung sind also nicht

Sie, sondern es ist das System. Im Folgenden werden Ihnen nacheinander [_*I1*_] Aufgaben zur Bearbeitung vorgelegt.

Lesen Sie bitte jedes Mal die Aufgabe laut vor, ehe Sie mit der Bearbeitung beginnen.

WICHTIG: Äußern Sie bei der Bearbeitung der Aufgaben Ihre Gedanken. Unterbrechen Sie dabei nicht die Bearbeitung der Aufgaben. Sagen Sie einfach alles, was Ihnen durch den Kopf geht. Dies können Gefühle, Gedanken, Absichten oder Erwartungen sein. Es gibt keine falschen Äußerungen.

Die Versuchsleitung wird sich Notizen machen, um Ihnen nach der Bearbeitung aller Aufgaben ggf. noch einige Fragen zu stellen.

[_*I2*_]

Selbstverständlich können Sie Ihre Teilnahme jederzeit und ohne Nachteile abbrechen. Wir möchten Sie allerdings bitten, alle Aufgaben bis zum Ende zu bearbeiten, da wir nur vollständige Datensätze für die Auswertung nutzen können. Die Untersuchung wird voraussichtlich [_*I3*_] dauern.

Zur Eingewöhnung beginnen wir mit einer Übungsaufgabe. Bitte lesen sie sie laut vor und beginnen Sie mit der Bearbeitung. Bitte vergessen Sie nicht, dabei laut zu denken.

[_*I4*_]

Vielen Dank für die Bearbeitung. Wenn Sie noch Fragen haben, stellen Sie sie bitte jetzt, da die Untersuchung möglichst nicht unterbrochen werden sollte.

[_*I5*_] [

Wir beginnen jetzt mit der eigentlichen Untersuchung. Bitte lesen Sie die erste Aufgabe laut vor. Beginnen Sie danach mit der Bearbeitung und denken Sie dabei laut.

Erläuterung:

[_*I1*_]: Anzahl der Aufgaben angeben.

[_*I2*_]: Ggf. ergänzen: Nach Abschluss der Aufgabenbearbeitung bitten wir sie abschließend einen Fragebogen auszufüllen.

[_*I3*_]: Dauer der Untersuchung angeben. WICHTIG: Realistische Schätzung der Dauer verwenden.

[_*I4*_]: Bearbeitung der Übungsaufgabe

[_*I5*_] Ggf. Beantwortung von Fragen; letzte Hinweise, falls etwas bei der Bearbeitung aufgefallen ist.

Anhang 2: Berechnung von Cohens Kappa

Die Berechnung von Cohens Kappa basiert auf einer **Kontingenztafel**, die die Häufigkeit übereinstimmender und divergierender Urteile repräsentiert. Im Falle von zwei Kodierungen ergibt sich eine 2x2 Matrix.

		Kodierung 1	
		1	0
Kodierung 2	1	a	b
	0	c	d

Zelle a enthält die Häufigkeit für Probleme, die in beiden Kodierungen enthalten sind, Zelle b die Häufigkeit von Problemen, die nur in der zweiten, nicht aber in der ersten Kodierung vorkommen, und Zelle c solche, die nur in der ersten, nicht aber in der zweiten Kodierung erscheinen. Da die Feststellung eines Usability-Problems auf Aussagen basiert, die als Indikator gewertet werden, enthält Zelle d die Häufigkeit von Aussagen, die in beiden Kodierungen ***nicht*** als Indikator für ein Problem klassifiziert werden. Die Häufigkeit N der Aussagen, die bei der Kodierung berücksichtigt werden, ergibt sich aus

$$N = a + b + c + d.$$

Ausgehend von der 2x2 Matrix, lässt sich Kappa nach folgender Formel berechnen:

$$K = (P_0 - P_e) / (1 - P_e)$$

mit:

$P_0 = (a + d) / N$

$P_e = ((a + b) / N * (a + c) / N) + ((c + d) / N * (b + d) / N)$

Zur Erläuterung sei Kappa für zwei Kodierungen von N = 500 Aussagen betrachtet, deren Beurteilung die 2x2 Matrix in nachfolgender Tabelle ergeben.

		Kodierung 1		Summe
		1	0	
Kodierung 2	1	50	10	60
	0	15	425	440
Summe		65	435	500

P_0 = (50 + 425) / 500 = 0,95

P_e = ((50 + 10 / 500 * (50 + 15) / 500) + ((15 + 425) / 500 * (10 + 425) / 500)

= (60 / 500 * 65 / 500) + (440 / 500 * 435 / 500)

= (0,12 * 0,13) + (0,88 * 0,87)

= 0,0156 + 0,7656

= 0,7812

K = (0,95 – 0,7812) / (1 – 0,7812)

= 0,1688 / 0,2188

= 0,7715

Online-Rechner zur Bestimmung von Kappa:

https://www.graphpad.com/quickcalcs/kappa1/?K=2

Anhang 3: Gliederungsbeispiel für Projektberichte zu Usability-Tests

1 Gegenstand des Tests

2 Zielsetzung und Fragestellung

3 Methode

- 3.1 Stichprobe
- 3.2 Messverfahren
- 3.3 Prüfaufgaben und Versuchsablauf

4 Ergebnisse

- 4.1 Konsolidierte Problemdokumentation
- 4.2 Erläuterungen zur Reliabilität
- 4.3 Quantifizierung und ggf. deskriptive Statistiken

5 Diskussion und Empfehlungen

- 5.1 Priorisierung der Problembehandlung auf Basis der Schweregrade
- 5.2 Optimierungsvorschläge

Anhang

Problemdokumentationen der Einzelkodierungen

Materialien