



ulm university universität  
**uulm**

Universität Ulm

Institut für Medieninformatik

Thorsten Mahler

Marc Hermann

Guido de Melo

Michael Weber

Studiengang Medieninformatik

Evaluation von User Interfaces

Medienpraktikum im Wintersemester 2007/2008

## **Ausarbeitung Evaluationsformen**

Eduard Seibel

Daniel Petrariu

Friedrich Hoermann

E-Mail:

[eduard.seibel@uni-ulm.de](mailto:eduard.seibel@uni-ulm.de), [daniel.petrariu@uni-ulm.de](mailto:daniel.petrariu@uni-ulm.de), [friedrich.hoermann@uni-ulm.de](mailto:friedrich.hoermann@uni-ulm.de)

Ausarbeitung Vortrag zu Evaluationsformen im Praktikum Evaluation von User-Interfaces

## **Begriffsklärung & Motivation**

Formal bedeutet der Begriff Evaluation das Beschreiben, Analysieren und Bewerten von Prozessen. Im Rahmen dieses Praktikums kann man diese Prozesse auf die Bedienung von User Interfaces näher eingrenzen. In diesem Sinne ist Evaluation das Testen des Designs eines Produktes, ob es so funktioniert wie gedacht und ob es den Anforderungen des Benutzers entspricht.

Die Anforderungen des Benutzers lassen sich in zwei Begriffen festhalten. Der Benutzer sollte durch das Produkt effektiv unterstützt werden, d. h. ihm soll das Produkt alle Funktionen bieten, welche zur Erledigung der Aufgabe(n) benötigt werden. Zudem sollte die Unterstützung effizient sein. In diesem Zusammenhang bedeutet Effizienz, dass die Aufgabe durch Verwendung des Produktes möglichst schnell erledigt werden kann.

Da die Benutzbarkeit maßgeblich zum Erfolg eines Produktes beiträgt, sollte Usability Engineering und die Evaluation als Teil davon während des gesamten Entwicklungsprozesses eines Produktes stattfinden. Die Evaluation ist also ein Mittel, um Probleme bei der Benutzung eines Produktes frühzeitig und systematisch aufzudecken, um diese anschließend beseitigen zu können. Frühzeitiges Aufdecken von Problemen in der Benutzung eines Produktes mittels Evaluation, führt i. d. R. zu beträchtlichen Zeit- und Geldersparnissen während des gesamten Entwicklungsprozesses. Aber auch bei fertigen Produkten können Probleme in der Benutzung mittels Evaluation gefunden werden und durch Beseitigung dieser Probleme das Produkt verbessert werden.

Im Folgenden sei deshalb auf die wichtigsten Arten und Methoden der Evaluation näher eingegangen.

## **Evaluationsarten**

Es gibt im Wesentlichen zwei Arten der Evaluation. Zum einen die Laborstudien und zum anderen die Feldstudien.

Bei Laborstudien findet das Testen eines Produktes in einer Testumgebung unter kontrollierten Bedingungen statt. Bei dem Testen von Software User Interfaces beispielsweise ist diese Testumgebung häufig ein neutraler Raum, in dem der Benutzer durch Video- und Audioaufzeichnungen während des Bedienvorgangs genau überwacht wird.

Die Vorteile dieser Art der Evaluation sind, dass der Benutzer unter kontrollierten Bedingungen störungsfrei arbeiten kann und dass i. d. R. bessere Untersuchungswerkzeuge

(z. B. Videokamera/Mikrofon) zur Verfügung stehen als bei einer Feldstudie. Dagegen ist Teamarbeit im Labor nur schwer simulierbar und die fehlende natürliche Arbeitsumgebung kann zu verändertem Nutzungsverhalten führen.

Laborstudien bieten sich deshalb vor allem dort an, wo der Arbeitsplatz gefährlich (Taucher) ist oder sehr weit entfernt liegt (z. B. Spaceshuttle). Weiterhin eignen sich Laborstudien zum Finden von speziellen Problemen durch gezielte Änderung des Kontextes oder Vergleich von unterschiedlichen Produktdesigns, deren Ergebnisse nur dann vergleichbar sind, wenn die Tests unter kontrollierten Bedingungen durchgeführt werden.

Bei Feldstudien wird der Benutzer in seiner normalen Arbeitsumgebung beobachtet und überwacht. Der Vorteil dieser Evaluationsart ist: dadurch dass sich der Benutzer in seiner natürlichen Arbeitsumgebung befindet, können aufgrund von Störfaktoren wie Lärm, Unterbrechungen etc. Beobachtungen auftreten, welche im Labor nicht gemacht werden können. Weiterhin können in Feldstudien langwierige Aufgaben beobachtet werden, in denen der Benutzer nicht für lange Zeit aus seiner Arbeitsumgebung herausgerissen werden kann.

Als nachteilhaft stellt sich bei Feldstudien heraus, dass Lärm, Ablenkung und viele weitere Störfaktoren die Ergebnisse von Studien mit veränderten Parametern nur schwer vergleichbar machen können und dass die eigentliche Arbeit durch die Studie unter Umständen unterbrochen werden muss.

Deshalb eignen sich Feldstudien vor allem dort, wo von einem starkem Einfluss der Arbeitsumgebung auf das Nutzungsverhalten auszugehen ist und bei Langzeitstudien.

### **Evaluation des Entwurfs**

Bei den nachfolgenden Methoden der Evaluation des Entwurfs werden die eigentlichen Benutzer nicht in die Evaluation mit einbezogen. Die Evaluation wird bei diesen Methoden durch den Entwickler selbst oder durch einen kognitiven Experten durchgeführt.

Bei der Methode des Cognitive Walktrough werden einzelne Aufgaben, welche Benutzer unter Einsatz eines Produktes durchführen, zunächst beschrieben und dann durch einen Experten evaluiert.

Dazu ist zunächst eine Prototypbeschreibung des Produkts nötig, welche durch ein Paper-Mockup geleistet werden kann. Dann werden einzelne Aufgaben, welche die Benutzer mit dem zukünftigen Produkt erledigen werden, extrahiert und z. B. durch GOMS beschrieben. Außerdem muss noch eine Beschreibung der Benutzerprofile erstellt werden, die

Informationen darüber liefert, was für Benutzer das Produkt später nutzen werden (erfahrene/unerfahrene Benutzer etc.). Sind alle Informationen zusammen, geht der Entwickler oder ein Experte jede Aufgabenbeschreibung sukzessive durch und stellt folgende Fragen:

1. Sind die einzelnen Schritte auch die, die man erwartet?
2. Gibt es Probleme beim Erlernen eines Schrittes?
3. Versteht der Benutzer die Antwort, die er vom System erhält?

Die Antworten auf diese Fragen werden dokumentiert und auf diese Weise nach und nach mögliche Probleme in der Benutzbarkeit eines Produktes aufgedeckt. Diese Methode ist sowohl für den frühen Entwurf, als auch für existierende Systeme geeignet. Nachteilhaft ist der hohe Zeitaufwand, welcher vor allem aus der Beschreibung der Aufgaben resultiert.

Bei der heuristischen Evaluation wird durch einen Satz von einfachen und allgemeinen heuristischen Regeln strukturiert Kritik am Produkt geübt. Der sicherlich bekannteste Satz von heuristischen Regeln sind die 9 Goldenen Regeln für eine gute Benutzerschnittstelle (gib ausreichend Feedback, minimiere die Gedächtnisleistung etc.). Diese Methode hat ein gutes Kosten-Nutzen Verhältnis, da viele grundlegende Fehler und Probleme durch diese Methode sehr schnell gefunden werden können.

Zu den Methoden die häufig in der Entwurfsphase zum Einsatz kommen, zählen auch verschiedene modellbasierte Evaluationen. Hier basiert die Evaluation auf einem Modell wie z. B. KLM (Keystroke Level Model) oder als Weiterentwicklung davon GOMS (Goals, Operators, Methods und Selection Rules).

Unter Einsatz der genannten Modelle können Aufgaben in ihre einzelnen Schritte zerlegt werden und basierend auf durchschnittlichen Bearbeitungszeiten für die einzelnen Schritte, wie z. B. das Positionieren des Mauszeigers an einer bestimmten Stelle oder das Drücken einer Taste, kann die gesamte Bearbeitungszeit einer bestimmten Aufgabe berechnet werden. In der Regel liefern solche Berechnungen gute Vergleichsschätzungen für zwei verschiedene Methoden, mit denen dieselbe Aufgabe erfüllt werden soll.

## **Evaluation der Implementierung**

Bei der Evaluation der Implementierung handelt es sich um die benutzerbasierten Evaluationsmethoden. Der Unterschied zur Evaluation des Entwurfs besteht darin, dass die späteren Benutzer des Systems mit in den Test einbezogen werden. Verschiedene Methoden dienen dieser Phase. Dies sind die kontrollierten Experimente, die Stumme Beobachtung, die Methode des lauten Denkens und die der konstruktiven Interaktion. Es gibt verschiedene Faktoren die wichtig sind für die Zuverlässigkeit des Experiments. Die sind zum einen die gewählten Testpersonen, die zu testenden Variablen und die getesteten Hypothesen. Die Testpersonen sollten so gewählt werden wie die spätere Nutzergruppe des zu evaluierenden Projekts sein wird. Sie sollten dieselbe Ausbildung und dasselbe Alter haben wie die Nutzer. Ihre Erfahrung beim Umgang mit ähnlichen Systemen, zum Beispiel Computern, sollte gleich sein. Um ein ausreichend aussagekräftiges Ergebnis zu erhalten, sollten mindestens 10 Testpersonen rekrutiert werden.

Bei den zu testenden Variablen gibt es zwei verschiedene Typen. Dies sind die unabhängigen, nicht messbaren und die abhängigen, messbaren Variablen. Unabhängige Variablen beschreiben beispielsweise das Interface Design eines Softwareprogramms, oder die Anzahl verschiedener Menüpunkte. Abhängige Variablen können eventuell die Geschwindigkeit bei der Menüauswahl oder die Fehlerrate sein.

Eine Hypothese ist eine Vorhersage des Ausgangs eines Experiments. Diese wird vor der Ausführung des Experiments aufgestellt und dann versucht während des Tests zu bestätigen. Die Vorhersage besagt, dass Änderungen unabhängiger Variablen die Werte der abhängigen Variablen beeinflussen. Normalerweise wird eine Nullhypothese aufgestellt, welche genau das Gegenteil der Hypothese besagt. Nun wird versucht diese Nullhypothese zu widerlegen. Dies wird gemacht, da die Falsifizierung einfacher ist, weil hierfür weniger Testergebnisse benötigt werden.

### **Ein Testdurchlauf könnte folgendermaßen aussehen:**

1. Hypothese aufstellen
2. Variablen spezifizieren
3. Testpersonen finden
4. Testverfahren wählen

- a. Zwischen-Gruppen-Test
  - b. In-Gruppen-Test
5. Auswertung/Analyse

Auch Mischungen aus Zwischen-Gruppen-Tests und In-Gruppen-Tests sind möglich und sogar oft erwünscht.

Bei Zwischen-Gruppen-Tests werden zwei Gruppen gebildet, die jeweils nur eine Variante des Projekts testen. Eventuell wird eine Gruppe als Kontrollgruppe benutzt um das Ergebnis zu bestätigen. Hierbei wird der Kontrollgruppe genau dieselbe Aufgabe gestellt wie einer der anderen Gruppen und danach das Ergebnis verglichen. Der Vorteil dieser Tests besteht darin, dass kein Lerneffekt zwischen den beiden Projektvarianten entsteht. Das heißt, Testpersonen lernen nicht bei der Benutzung einer der Varianten die Bedienung und können diese dann bei einem eventuellen zweiten Test besser oder schneller. Dies könnte zu Verzerrungen bei der Testdurchführung führen. Nachteil der Zwischen-Gruppen-Tests ist jedoch die höhere Anzahl der benötigten Testpersonen.

Bei In-Gruppen-Tests sind alle Testpersonen in derselben Gruppe und testen verschiedene Varianten eines Projekts. Vorteil dieser Methode ist die geringere Anzahl benötigter Probanden. Nachteil ist natürlich der bereits oben angesprochene Lerneffekt von einer Variante zur anderen.

Die Analysephase sei hier nur kurz beschrieben, genauere Beschreibungen finden sich in anderen Ausarbeitungen. Es ist wichtig, dass während des Tests alles gespeichert und aufgezeichnet wird. Dazu gehören auch eventuelle Störungen und Unterbrechungen während des Tests. Die Messdaten sollten auch immer von Hand durchgegangen werden um eventuelle Ausreißer in den Daten zu finden und diese extra zu behandeln. Beispielsweise benötigt eine Person eventuell die fünffache Zeit für eine Aufgabe, weil sie weniger Erfahrung mit dem zu evaluierenden System oder ähnlichen System hat als andere Testpersonen.

Nachfolgend nun ein kleines Beispiexperiment, um den groben Ablauf zu zeigen:

Das zu testende System sei ein einfacher Stadtplan mit unterschiedlichen Buttons als Menüleiste, um den Plan zu zoomen und zu verschieben. Das System soll nach der Fertigstellung für Touristen in der Ulmer Innenstadt zur Verfügung stehen. Diese Buttons sollen nun zur schnelleren Bedienung unterschiedlich eingefärbt werden.

**Frage:** Kann das Menü des Testsystems schneller bedient werden, wenn die Icons eingefärbt werden?

**Testpersonen:** Beliebig aus der Bevölkerung

**Hypothese:** Einfärbung wird die Bediengeschwindigkeit erhöhen.

**Nullhypothese:** Es wird keinen Unterschied zwischen normalem und eingefärbtem Menü geben.

**Unabhängige Variablen:** Die Einfärbung.

**Abhängige Variablen:** Bediengeschwindigkeit gemessen in Sekunden.

**Testdesign:** Zwei Gruppen Test, je eine mit Version ohne Farbe und eine mit.

**Analyse:** t-Test

### **Beobachtungsmethoden**

Bei der stummen Beobachtung bekommt der Benutzer eine Aufgabe und wird vom Tester beobachtet. Es gibt keine Kommunikation zwischen der Testperson und Person, welche die Evaluation durchführt. Nachteil dieser Methode ist, dass der Tester keinerlei Einblick in die Denk- und Entscheidungsprozesse des Benutzers bekommt.

Besser ist hier die Methode des lauten Denkens (Think Aloud). Dabei wird der Benutzer aufgefordert zu sagen was er gerade denkt. Dies bringt leicht gröbere Interface Probleme zu Tage. Nachteile können sein, dass es unangenehm und ungewohnt für die Testpersonen ist. Lautes Denken ist unnatürlich und kann dadurch das normale Verhalten beeinflussen. Weiter kann es sein, dass der Benutzer während des Tests zu konzentriert sein muss um nebenbei zu reden.

Eine weitere Variante ist die konstruktive Interaktion, wobei zwei Personen gemeinsam an einer Aufgabe arbeiten. Dadurch fühlt sich die Testperson nicht so sehr als Versuchsperson, sondern eher als normaler Mitarbeiter. Vorteil gegenüber der Methode des lauten Denkens ist hierbei, dass der Vorgang leichter zu lernen ist für die Testnutzer. Der Benutzer kann auf problematische Stellen hinweisen zu genau dem Zeitpunkt, an dem sie auftreten. Dies maximiert die Effektivität um Problemstellen zu finden.

### **Wizard of Oz**

Bei der Wizard of Oz Methode wird dem Testbenutzer vorgespielt, dass er mit dem System kommuniziert. In Wirklichkeit erzeugt jedoch einer der Tester im Verborgenen die Reaktion des Systems. Solche Experimente werden durchgeführt, um mögliche Reaktionen von potenziellen Benutzern eines Systems zu sammeln das gerade erst entwickelt wird.

## **Aufzeichnungsmethoden**

Die Aufzeichnung mit Stift und Papier ist billig, jedoch begrenzt durch die Schreibgeschwindigkeit des Testers.

Die Audioaufzeichnung ist gut für die Methode des lauten Denkens, jedoch ist es oftmals schwer eine Audiostelle einer durchgeführten Aktion zuzuordnen.

Bei der Videoaufzeichnung ist der Vorteil, dass man sieht was der Benutzer genau macht. Dabei sind zwei Kameras sehr nützlich, wobei dabei eine auf den Bildschirm gerichtet wird und eine auf den Benutzer. Allerdings kann diese Methode auch als sehr aufdringlich empfunden werden und somit das natürliche Verhalten beeinflussen.

Oft ist ein Mix aus allem am besten geeignet, beispielsweise Videoaufzeichnungen plus kurze Notizen von Hand. Für die Auswertung von Videoaufzeichnungen sollte genügend Zeit eingeplant werden. Man rechnet im Durchschnitt mit zwei Stunden Auswertungszeit für eine Stunde aufgezeichnetes Material.

## **Nach-Test Durchlauf**

Manchmal kommt es vor, dass selbst bei der Methode des lauten Denkens nicht klar ist, warum ein Benutzer eine bestimmte Aktion durchgeführt hat. Hier empfiehlt es sich die Aufzeichnungen nochmals zusammen mit der Testperson durchzugehen. Dies kann entweder direkt nach dem Test geschehen oder zeitlich versetzt. Vorteil des direkten Nach-Test Durchlaufs ist die gute Erinnerung bei der Testperson an den Test. Vorteil des zeitlich versetzten Nachtests ist die Möglichkeit konkretere Fragen vorzubereiten.

## **Wichtiges für den Test**

Beim Test selbst sollte keine Zeit der Testpersonen verschwendet werden und die Tester sollten sich wohl fühlen. Hierzu könnte man je nach Budget für eine kleine Aufwandsentschädigung sorgen. Bei langen Tests sind kurze Pausen wichtig und auch Kleinigkeiten zu Essen sollten angeboten werden. Unterbrechungen während des Tests sollten so gut wie möglich vermieden werden. Vor dem Test sollte man den Benutzern Vertraulichkeit garantieren und nur die absolut nötigsten Dinge zur Durchführung des Tests erklären. Dies könnte sonst die Ergebnisse ebenfalls stark beeinflussen.