Usability Engineering

Vorlesung 6: Usability Testing (Teil 3)

VU 183.123

Christoph Wimmer

deco
designing comfort

http://deco.inso.tuwien.ac.at

Übersicht

- Usability Testing:
  - Vorbereitung & Planung
  - Durchführung
    - Ablauf der Test-Sessions
    - Guidelines für Teammitglieder
    - Datenerfassung
  - Auswertung
  - Kommunikation der Ergebnisse

Auswertung

Auswertung

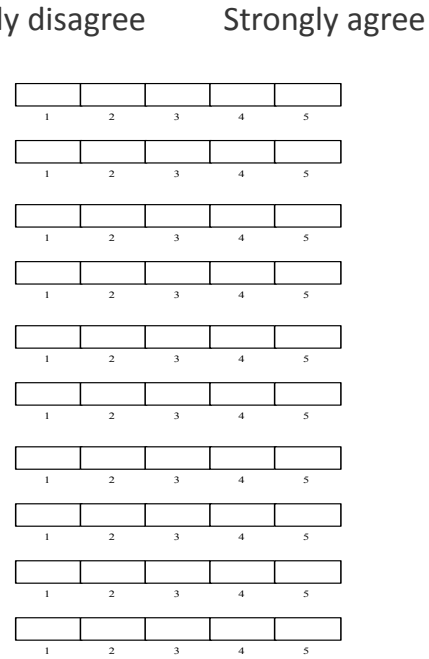
- Die unterschiedlichen Informationen werden gesammelt
  - Unterschiedliche Quellen (Notizen von unterschiedlichen Beobachtern, Audio, Video, Logfiles, erhobene Messwerte, …)
- Die Beobachtungen werden geordnet
- Die Beobachtungen werden analysiert
  - Wie lassen sich die Beobachtungen zu Problemen zusammenfassen?
  - Welche Probleme traten (häufig) auf?
  - Zu welcher Kategorie gehört das Problem?
  - Was ist eine mögliche Ursache des Problems?
  - Wie könnte man das Problem lösen?

Auswertung

- Quantitative Daten – Wie viele & wie oft:
  - Anzahl der Testpersonen, die erfolgreich waren
  - Benötigte Zeit
  - Anzahl der Fehler (**je Task**, evtl. je Person)
  - Anzahl der Hilfestellungen (**je Task**, evtl. je Person)
  - Häufigkeit von Problemen
  - Anzahl der Operationen
  - …
- Deskriptive Statistik (z.B. um einen Gesamteindruck zu gewinnen)
- Signifikanztests (zum Vergleich von unterschiedlichen Alternativen)
- Zur Abschätzung des Schweregrads, als Argumentationsgrundlage und um ein Gesamtbild zu vermitteln

Auswertung: Beispiel SUS

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

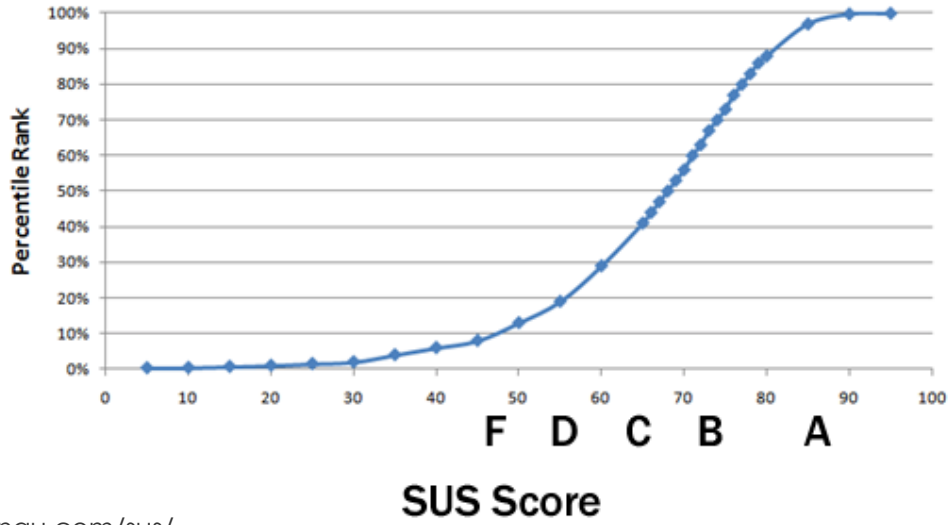


Auswertung: Beispiel SUS

- 10 Fragen auf einer 5-Punkte Likert-Scale zu beantworten
  - Frage 1,3,5,7,9: Positiv formuliert
  - Frage 2,4,6,8,10: Negativ formuliert
- Scoring:
  - Skalierung auf Werte von 0 - 4 (4 = positiv, 0 = negativ)
    - Bei positiv formulierten Fragen: Antwort-Wert - 1
    - Bei negativ formulierten Fragen: 5 - Antwort-Wert
  - Skalierte Werte addieren und mit 2,5 multiplizieren
  - Ergebnis: Score zwischen 0 (= negativ) und 100 (= positiv)

Auswertung: Beispiel SUS

- Ergebnis: Score zwischen 0 (= negativ) und 100 (= positiv)
- Achtung: Dieser Wert ist kein Prozentwert **(!)**
  - Diverse Studien haben gezeigt, dass der Median SUS-Score bei ca. 70 liegt
  - 90 % Quantil liegt bei einem SUS-Score von ca. 80



Siehe auch: https://measuringu.com/sus/

Auswertung

- Qualitative Daten – Wie & wieso:
  - Aussagen und Verhaltensweisen
  - Subjektive Meinungen
  - Muster und Auffälligkeiten
  - Interpretation der BeobachterInnen

- Gruppierung nach Gemeinsamkeiten
- Klassifikation
  - Beobachtung oder Interpretation?
  - Art des Problems?
  - Ursache?

Von Beobachtungen zu Usability Problemen

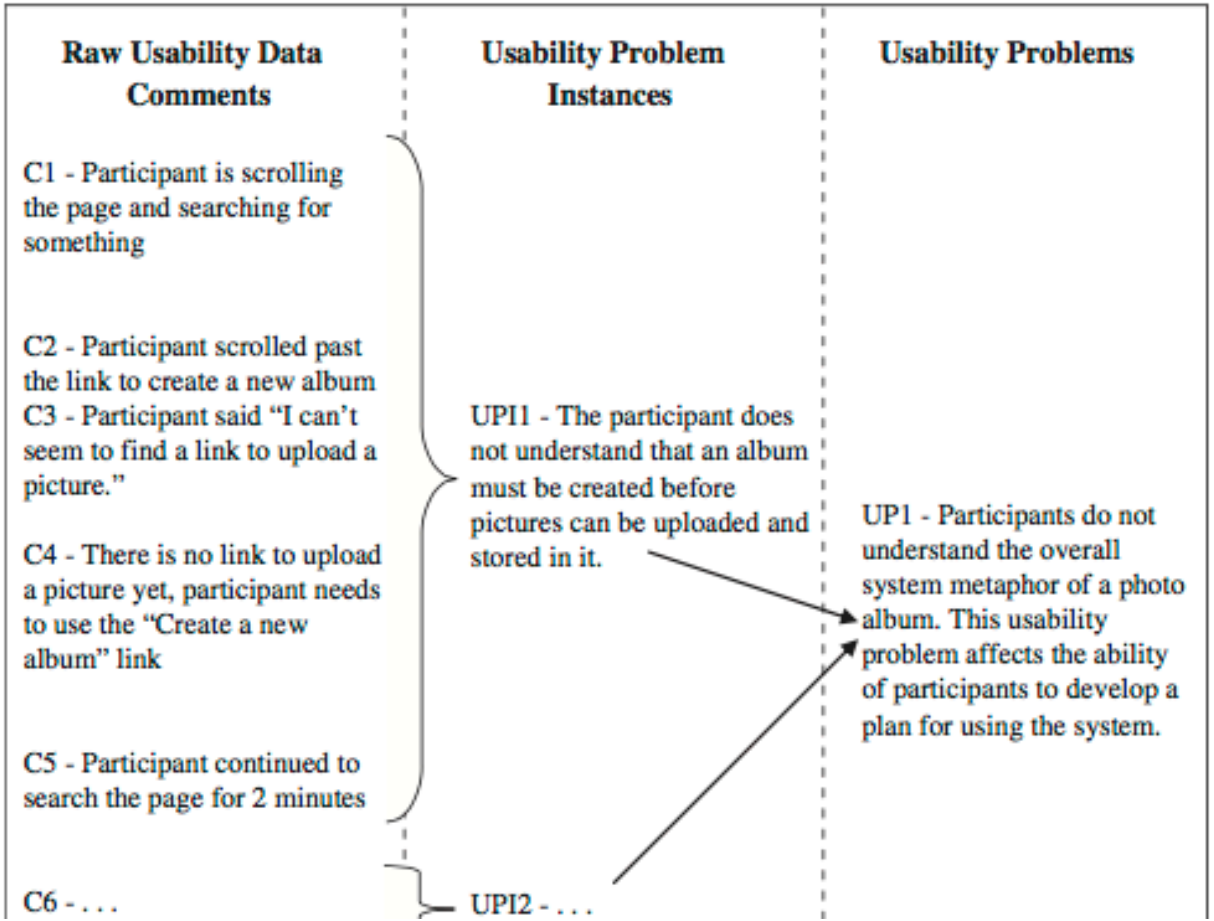
- Einzelne, individuelle **Beobachtungen** aus den Sessions (Rohdaten) werden zusammengefasst zu

➔ **Usability Probleminstanzen**, welche die individuellen Beobachtungen zusammengefasst beschreiben, werden zusammengefasst zu

➔ **Usability Problemen**, welche Auslöser und Ursachen des Problems beschreiben

- Problem Matching z.B. über **Similar Changes**: Zwei Probleminstanzen werden zusammengefasst, wenn mit der Behebung der einen Probleminstanz mit großer Wahrscheinlichkeit auch die andere behoben wird.

Von Beobachtungen zu Usability Problemen



Siehe auch: Howarth et al. (2009), Supporting novice usability practitioners with usability engineering tools

Problembeschreibung

- Be clear and precise while avoiding wordiness and jargon
  - Be concrete, not vague
  - Be practical, not theoretical
  - Use descriptions that non-HCI people will appreciate
  - Avoid so much detail that no one will want to read the description
- Describe the impact and severity of the problem
  - Describe how it impacts the user's task
  - Describe how often the problem will occur, and system components that are affected or involved

Problembeschreibung

- Support your findings with data
  - Include information on how many users experienced the problem and how often
  - Include objective data, both quantitative and qualitative, such as the number of times a task was attempted or the time spent on the task
  - Provide traceability of the problem to observed data
- Describe the cause of the problem
  - Describe the main usability issue involved in the problem
  - Avoid guessing about the problem cause or user's thoughts

Problembeschreibung

- Describe observed user actions
  - Include contextual information about the user and the task
  - Include specific examples, such as the user's navigation flow through the system, user's subjective reactions, screen shots, and task success/failure
  - Mention whether the problem was user reported or experimenter observed
- Describe a solution to the problem
  - Provide alternatives and tradeoffs
  - Be specific enough to be helpful without dictating a solution
  - Supplement with usability design principles

Usability Probleme: Severity Rating

- Severity Rating = Gewichtung, Priorisierung

- Wird benutzt für:
  - Einteilung der Ressourcen für die Behebung des Problems
  - Schätzung, ob weitere Usability Engineering Maßnahmen notwendig sind
- Beurteilung aus Kombination von:
  - Häufigkeit: Tritt das Problem häufig oder selten auf?
  - Persistenz: Tritt es einmal oder immer wieder (auch wenn man es als BenutzerIn einmal überwunden hat) auf?
  - Bedeutung für den/die BenutzerIn: Wie schwerwiegend sind die Auswirkungen des Problems?

Schlussfolgerungen aus den Ergebnissen

- Direkt mit den Testzielen vergleichen:
  - Wurden die Testziele erreicht? Wurden die Fragen beantwortet?
  - Warum wurden Sie nicht erreicht bzw. beantwortet?

- Zu den High-Level Zielen in Beziehung setzen:
  - einfach zu benutzen
  - effizient zu benutzen
  - einfach zu erlernen
  - zufriedenstellend

- Verbesserungsvorschläge zur Verbesserung der Applikation ergeben sich aus den Details der Ergebnisse und nicht aus der Zusammenfassung
  - Wieso war BenutzerIn XY verwirrt?
  - Wieso war die Interaktion schwierig?

Schlussfolgerungen aus den Ergebnissen

- Unterschiedliche Ursachen für dasselbe Problem möglich
- Beispiel: Testpersonen finden einen Befehl in einem Menü nicht
- Mögliche Ursachen:
  - Die Benennung des Befehls ist un-/mißverständlich
  - ➡ Möglicher Lösungsansatz: Befehl umbenennen
  - Der Befehl wird an anderer Stelle in der Menüstruktur erwartet
  - ➡ Möglicher Lösungsansatz: Befehl verschieben
  - Die Menüstruktur als Ganzes ist unübersichtlich bzw. schlecht strukturiert, …
  - ➡ Möglicher Lösungsansatz: Umfassendes Redesign der Menüstruktur
  - Die Testpersonen wissen nicht, dass es den Befehl überhaupt gibt
  - ➡ Möglicher Lösungsansatz: ?
  - …

➡ Unterschiedliche Ursachen erfordern unterschiedliche Lösungsansätze

Kommunikation der Ergebnisse

Kommunikation der Ergebnisse

- Direkt: Beobachtung vor Ort
- Schriftlich: In Form eines Berichts
  - Dient als Dokumentation für später
  - Auf Basis des Berichts sollte man die genaue Vorgehensweise rekonstruieren können
- Verbal: In Form eines Vortrags, einer Präsentation oder einem Meeting
- Visuell: In Form eines Videos
  - Als zusätzliches Mittel, um die gefundenen Probleme zu verdeutlichen (vorausgesetzt es wurden Videos der Tests erstellt)

Aufbau Testbericht

Management Summary

Problemstellung (Produktbeschreibung, Testziele)

Methode (Testpersonen, Testszenarien, Testmaterial und Einrichtung, Design, Ablauf, generelle bzw. taskspezifische Anweisungen an die Testpersonen)

Ergebnisse (nach Testszenario gruppiert)

Analyse und Verbesserungsvorschläge

Conclusio

Anhänge (Interviewleitfaden, Fragebögen, Testmaterial, etc.)

Vergleiche: Common Industry Format (CIF) for Usability Test Reports, https://www.nist.gov/publications/common-industry-format-usability-test-reports

Aufbau Testbericht: Vor dem Test

Management Summary

Problemstellung (Produktbeschreibung, Testziele)

Methode (Testpersonen, Testszenarien, Testmaterial und Einrichtung, Design, Ablauf, generelle bzw. taskspezifische Anweisungen an die Testpersonen)

Ergebnisse (nach Testszenario gruppiert)

Analyse und Verbesserungsvorschläge

Conclusio

Anhänge (Interviewleitfaden, Fragebögen, Testmaterial, etc.)

Vergleiche: Common Industry Format (CIF) for Usability Test Reports, https://www.nist.gov/publications/common-industry-format-usability-test-reports

Aufbau Testbericht: Nach dem Test

Management Summary

Problemstellung (Produktbeschreibung, Testziele)

Methode (Testpersonen, Testszenarien, Testmaterial und Einrichtung, Design, Ablauf, generelle bzw. taskspezifische Anweisungen an die Testpersonen)

Ergebnisse (nach Testszenario gruppiert)

Analyse und Verbesserungsvorschläge

Conclusio

Anhänge (Interviewleitfaden, Fragebögen, Testmaterial, etc.)

Vergleiche: Common Industry Format (CIF) for Usability Test Reports, https://www.nist.gov/publications/common-industry-format-usability-test-reports

Bericht – Tips zur Ausarbeitung

- Kommentare & Zitate der Testpersonen einbauen
- Beispiele (inkl. Screenshots)
- „Alle 10 Testpersonen scheiterten …“
- „8 von 10 hatten keine Probleme …, aber 2 Testpersonen …“
  - Nicht „80% der Testpersonen …“
- Pro Task zusammenfassen
- Probleme bewerten – Severity Rating
- Positive Ergebnisse einbeziehen
- Verbesserungsvorschläge anbieten

Zusammenfassend...

- **Ein** Test ist besser als **kein** Test
  - Usability Testing hat (afaik) noch kein Produkt schlechter gemacht
  - Fast jeder kann einen Usability Test durchführen, man muss es nur machen
- Man testet, um etwas zu **lernen** und **Fehler zu finden**...
  - ... und meistens findet man auch welche
- Usability Testing kann (und sollte?) auch **Spaß** machen

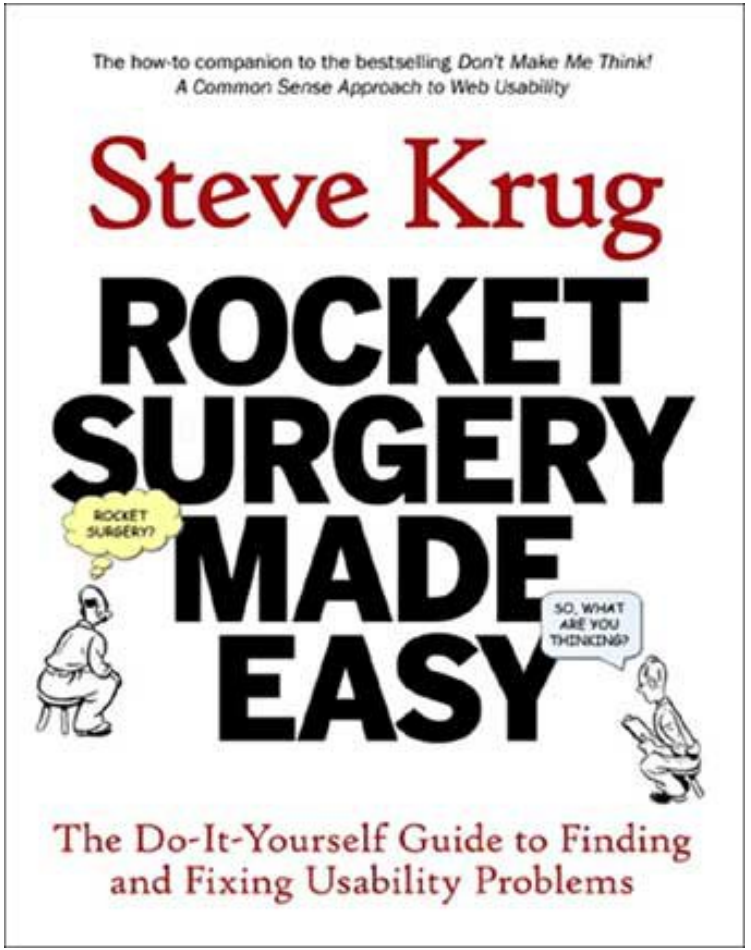
Buchtip: Rocket Surgery Made Easy

- Steve Krug, **Rocket Surgery Made Easy**

  (New Riders, 2010)

- Begleitmaterial:

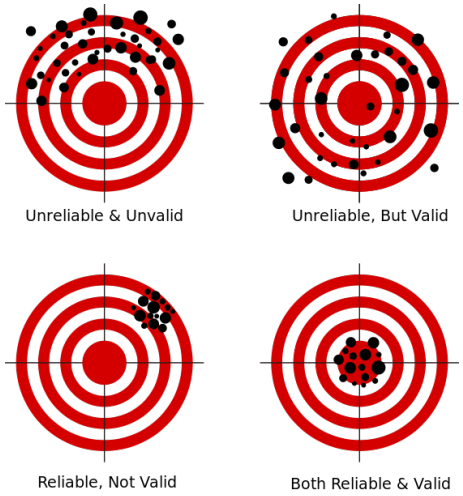
  http://www.sensible.com/rsme.html



Qualität von Evaluierungsmethoden

Qualität der Methode

- Vollständigkeit: Finde ich möglichst viele Probleme?
- Objektivität: Sind meine Ergebnisse von subjektiven Einflüssen der durchführenden Personen unbeeinflusst?
- Reliabilität: Sind meine Ergebnisse konsistent? Lassen sich die Ergebnisse unabhängig reproduzieren?
- Validität: Sind meine Ergebnisse korrekt?
  - Intern: Vermeidung von Störvariablen
  - Extern: Generalisierbarkeit über den Test hinaus



Qualität der Methode

- Gesamtqualität in der Praxis schwer exakt festzustellen
  - Vollständigkeit: Es gibt keine Möglichkeit mit Sicherheit festzustellen, dass man alle Probleme gefunden hat, weil es keine perfekte Usability und keine perfekte Evaluierungsmethode gibt
  - Schwere der Probleme: Wird in den meisten Studien zur Effektivität von Methoden kaum berücksichtigt
  - Objektivität: Klassifikation und Analyse qualitativer Daten erlaubt immer einen gewissen Interpretationsspielraum (Subjektivität)
  - Reliabilität: In der Praxis werden Usability Tests nur selten wiederholt

- Dennoch: Fallstudien und Meta-Studien geben einen Überblick über Effektivität unterschiedlicher Methoden

Usability Testing

If you ask ten usability experts how to run a valid usability test, you'll get about twenty different answers.

- Lukas Mathis

Comparative Usability Evaluation

- Das Test-Design für einen Usability Test bietet viele Gestaltungsmöglichkeiten
- Wie können diese Gestaltungsmöglichkeiten die Ergebnisse eines Usability Tests beeinflussen?

- CUE Website: http://www.dialogdesign.dk/CUE.html
- 10 Studien zur gängigen Praxis und Reproduzierbarkeit der Usability Evaluierung

Team	A	B	C	D
1. Number of reported problems	4	98	25	35
2. Number of reported problems that include specific recommendations for improving the interface	0	24	6	35
3. Number of reported problems that were encountered by one user only	0	4	2	8
4. Number of reported problems that deal exclusively with aesthetics (choice of colors, etc.)	0	0	5	1
5. Problems classified by severity Recommended in [1]	All four problems are severe	No	No	No
6. Number of positive findings reported. Recommended in [1]	1	4	3	0
7. Number of reported suggestions from test participants for improving the interface	0	2	5	0
8. Number of program errors reported	0	1	0	0
9. Indication of how many users encountered each problem Recommended in [1]	Yes	No	No	No

Team	A	B	C	D
1. Number of reported problems	4	98	25	35
2. Number of reported problems that include specific recommendations for improving the interface	0	24	6	35
3. Number of reported problems that were encountered by one user only	0	4	2	8
4. Number of reported problems that deal exclusively with aesthetics (choice of colors, etc.)	0	0	5	1
5. Problems classified by severity Recommended in [1]	All four problems are severe	No	No	No
6. Number of positive findings reported. Recommended in [1]	1	4	3	0
7. Number of reported suggestions from test participants for improving the interface	0	2	5	0
8. Number of program errors reported	0	1	0	0
9. Indication of how many users encountered each problem Recommended in [1]	Yes	No	No	No

Siehe auch: Molich et al. (1999), Comparative Evaluation of Usability Tests

Team	A	B	C	D
1. Total person hours used for the test by the usability professionals. Test participants' time is not included. Equal to the sum of the following rows 2-4.	26	70	24	84
2. Time used for planning and usability context analysis	9	10	6	28
3. Time used for recruiting test participants and testing. Test participants' time is not included.	12	20	8	21
4. Time used for analysis of results and reporting	5	40	10	35
5. Number of usability professionals involved	2	2	1	3
6. Number of tests	18	5	4	5
7. Approximate length of each usability test in minutes	4 to 32	120	120	60
8. Profiles of test participants reported	No	No	Yes	Yes
9. Number of scenarios/tasks used in test	5	11	5	4
10. Detailed scenario descriptions provided (see also table 5)	Yes	Yes	Partly	Yes
11. Quantitative assessment of user interface provided	Yes	No	No	Yes
12. Results of heuristic evaluation performed by usability professional included in report	No	No	Yes	No

Team	A	B	C	D
1. Total person hours used for the test by the usability professionals. Test participants' time is not included. Equal to the sum of the following rows 2-4.	26	70	24	84
2. Time used for planning and usability context analysis	9	10	6	28
3. Time used for recruiting test participants and testing. Test participants' time is not included.	12	20	8	21
4. Time used for analysis of results and reporting	5	40	10	35
5. Number of usability professionals involved	2	2	1	3
6. Number of tests	18	5	4	5
7. Approximate length of each usability test in minutes	4 to 32	120	120	60
8. Profiles of test participants reported	No	No	Yes	Yes
9. Number of scenarios/tasks used in test	5	11	5	4
10. Detailed scenario descriptions provided (see also table 5)	Yes	Yes	Partly	Yes
11. Quantitative assessment of user interface provided	Yes	No	No	Yes
12. Results of heuristic evaluation performed by usability professional included in report	No	No	Yes	No

CUE: 5 Key Findings

- **Five users are not enough:** Five users are enough to drive a useful iterative cycle, but never claim that you found all usability problems in an interactive system.
- **Huge number of issues:** The total number of usability issues for the state-of-the-art websites that we have tested is huge, more than 300 and counting. It is much larger than you can hope to find in one usability test.
- **Usability inspections are useful:** The CUE-4 study indicated that usability inspections produce results of a quality comparable to usability tests—at least when carried out by experts.
- **Designing good usability test tasks is challenging:** In CUE-2, nine teams created 51 different tasks for the same user interface. We found each task to be well designed and valid, but there was scant agreement on which tasks were critical.
- **Quality problems in some usability test reports:** The quality of the usability test reports varied dramatically. Some reports lacked positive findings, executive summaries, and screen shots. Others were complete with detailed descriptions of the team's methods and definitions of terminology.

Inspektionsmethoden vs. Usability Test

Karat et al. (1992): Comparison of Empirical Testing and Walkthrough Methods in User Interface Evaluation

Inspektionsmethode vs. Usability Test:

- Einzel-Walkthrough: 142 Problem Tokens, 35 SPAs, 55 NAAs, 2 UPAs
- Team Walkthrough: 222 Problem Tokens, 37 SPAs, 53 NAAs, 1 UPA
- Usability Test: 822 Problem Tokens, 67 SPAs, 23 NAAs, 21 UPAs

- 80 SPAs ingesamt
- 23 SPAs mit allen 3 Methoden gefunden

SPA = Significant Problem Area

NAA = No Action Area

UPA = Unique Usability Problem Area

deco.inso.tuwien.ac.at