# Usability Test Ergebnisse – Eine sehr persönliche Angelegenheit

**Molich, Rolf**
DialogDesign
Skovkrogen 3
DK-3660 Stenløse
Dänemark
Molich@DialogDesign.dk

**Daske, Lisa**
Astrum IT GmbH
Am Wolfsmantel 2
91052 Erlangen
Deutschland
post@lisadaske.de

**Abstract**
ADer Beitrag stellt einige der wichtigsten Ergebnisse der Usability Test Studie CUE-9 vor. In CUE-9 (Comparative Usability Evaluation-9) erhielten 35 Usability Professionals die selben 5 Usability Test Videos von einer Webseite zur unabhängigen Auswertung. Später haben die Teilnehmer ihre Auswertungen verglichen, gestaunt, und von den großen Unterschieden gelernt.
Wichtige Ergebnisse sind: Auf der getesteten Webseite einer Autovermietungsfirma wurden mehr als 200 Usability Probleme gefunden.Die Teilnehmer fanden größtenteils unterschiedliche Probleme. Nur wenige Probleme wurden von mehr als der Hälfte der Teilnehmer gefunden. Der Schweregrad von Problemen wurde oft sehr unterschiedlich beurteilt.

**Keywords:**
/// CUE-9
/// Usability Testing
/// Thinking-Aloud-Methode
/// Comparative Usability Evaluation

## 1. Einleitung

Um die Usability von Systemen zu verbessern, muss die Usability dieser Systeme zuerst bewertet werden können. Dafür werden verlässliche und robuste Evaluierungsmethoden benötigt.

Der Usability Test wird von manchen für die wichtigste aller Evaluierungsmethoden gehalten und er wurde bereits öfter als Maßstab für andere Evaluierungsmethoden verwendet. Diese prominente Stellung des Usability Tests ist eine Motivation dafür, die Methode genau zu untersuchen, um ihre Stärken und Schwächen kennenzulernen und die Methode innerhalb ihrer Grenzen einzusetzen. Das ist die Motivation hinter Rolf Molichs Studienreihe CUE („Comparative Usability Evaluation“, zu deutsch „Vergleichende Usability Evaluierung“).

Bei jeder CUE-Studie evaluieren Teams aus Usability Professionals unabhängig voneinander die selbe Webseite, Web-Anwendung oder Software. Das Hauptziel der Studien ist es, genug Daten zu sammeln, um unter anderem folgende Fragen zu beantworten:

- Ist das Ergebnis von Usability Evaluierungen reproduzierbar?
- Wie viele Usability Probleme hat das getestete Produkt tatsächlich?
- Wie viele Usability Probleme findet man auf einer typischen, nicht trivialen Webseite normalerweise?
- Was ist ein „kritisches“ oder „ernstes“ Usability Problem?

Die neunte Studie in dieser Reihe, CUE-9, wurde im Rahmen der UPA-I Konferenz in Atlanta (USA) 2011 sowie im Rahmen der „Mensch und Computer“-Konferenz in Chemnitz 2011 mit insgesamt 35 Usability Professionals durchgeführt. Dieser Beitrag stellt einige der wichtigsten Ergebnisse der CUE-9-Studie vor.

## 2. Ablauf der Studie

Der Fokus der Studie war es, herauszufinden, ob ein „Evaluator Effect“ existiert. Dieser besagt, dass auf der Grundlage der selben Daten mehrere Experten unterschiedliche Usability Probleme im selben System finden. Mehr Informationen zu CUE-Reihe finden Sie auf der Webseite von DialogDesign (1).

Bei einem gemeinsamen Workshop trafen sich die Studienteilnehmer (getrennt in Deutschland und Amerika), um ihre Ergebnisse zu vergleichen und zu diskutieren. Diese Workshops fanden am 20. Juni 2011 in Atlanta, GA, USA statt (CUE-9a), sowie am 11. September 2011 in Chemnitz, Deutschland (CUE-9b). In Atlanta nahmen dabei 19 Usability Experten teil, in Chemnitz waren es 16 Teilnehmer.

### 2.1. Vorbereitung

Jeder der Teilnehmer erhielt im Vorfeld 5 Videos von Usability Test Sitzungen zur unabhängigen Auswertung. Grundlage der Studie war eine gemeinsame Evaluierung der Webseite der US-amerikanischen Firma U-Haul, www.u-haul.com, die LKWs an Privatpersonen zu Umzugszwecken vermietet. Den Studienteilnehmern wurden folgende Aufgaben gestellt:

- Fünf Videos zu je 30 Minuten aus Usability Tests ansehen und bewerten
- Einen kurzen, anonymen Testbericht schreiben und einreichen
- Ähnliche Testberichte von anderen Studienteilnehmern lesen

### 2.2. Workshop

Während der Workshops wurden die Teilnehmer in Gruppen von 4–5 Personen geteilt. Die Gruppen erhielten einzelne Usability Probleme in gedruckter Form, die:

- Von den Teilnehmern der Gruppe in deren jeweils einzelner Evaluation gefunden worden waren
- In den einzelnen Evaluationen als katastrophal, kritisch, ernst, oder als Bug bewertet wurden (AA, A, B oder X auf den Bewertungsskalen)

Die Gruppen wurden gebeten, die Ergebnisse zu einer Gruppenevaluation zusammenzuführen. Im zweiten Teil des Workshops diskutierten die Teilnehmer über ihre Eindrücke. Gegenstand der Diskussion war, ob die Gruppenevaluation neue Erkenntnisse beigetragen hatte, und ob die Teilnehmer einen „Evaluator Effect" wahrgenommen hatten.

## 3. Ergebnisse

### 3.1. Evaluator Effect

Es wird angenommen, dass der Hauptgrund für den Evaluator Effect der ist, dass die Usability Evaluierung eine interpretierende Tätigkeit ist. Die Tester müssen Ermessensentscheidungen treffen, indem sie von einer Abfolge von Benutzerinteraktionen auf eine von Usability Problemen schließen. Es ist nicht überraschend, dass solche Entscheidungen bei verschiedenen Testern nicht immer zum selben Ergebnis führen. Was überraschen kann, ist das Ausmaß des Evaluator Effects.

In vorherigen CUE-Studien war die Anzahl der Probleme, die nur von einem einzigen Tester gefunden wurden, deutlich größer als die Anzahl der Probleme, die von mehreren Testern gefunden wurden. Auch in CUE-9 konnte dieser Effekt gezeigt werden: Immerhin 37% der Probleme wurden nur von einem einzigen der Teilnehmer berichtet. Der Anteil derjenigen Probleme, die von 4 oder mehr Teilnehmern berichtet wurden, beträgt ebenfalls lediglich 38%. **[Tab. 1]**

| | CUE-9a | CUE-9b | Combined |
|---|---|---|---|
| Anzahl aller Probleme | 134 | 93 | 169 |
| Probleme, die nur ein Teilnehmer berichtete | 52/134<br>39% | 35/93<br>38% | 62/169<br>37% |
| Probleme, die von 2 oder mehr Teilnehmern berichtet wurden | 82/134<br>61% | 58/93<br>62% | 107/169<br>63% |
| Probleme, die von 4 oder mehr Teilnehmern berichtet wurden | 48/134<br>36% | 32/93<br>34% | 65/169<br>38% |

**Tab. 1.**
Anzahl der von mehreren Teilnehmern berichteten Probleme

| | CUE-9a | CUE-9b | Combined |
|---|---|---|---|
| Anzahl Prüfer | 19 | 16 | 35 |
| Anzahl aller Befunde | 860 | 473 | 1,333 |
| Anzahl Befunde nach Kombination ähnlicher Befunde | 182 | 109 | 222 |
| Anzahl positiver Befunde | 48 | 16 | 53 |
| Anzahl negativer Befunde (Probleme) | 134 | 93 | 169 |

**Tab. 2.**
Anzahl Befunde

#### 3.1.1. Wahrnehmung der Teilnehmer

In Atlanta gaben 13 von 19 Teilnehmern an, kritische und ernste Probleme übersehen zu haben, die von anderen Teilnehmern in ihren Gruppen erkannt wurden. Dass die Mehrzahl der Teilnehmer einen Evaluator Effect wahrgenommen hatten, wurde in der Plenumsdiskussion klar ersichtlich. So sagte ein Teilnehmer: „Ich kam in diesen [Workshop] und dachte, es müsste mehr Übereinstimmung geben als es tatsächlich gab. (…)" Ein anderer Teilnehmer drückte es so aus: „Ich finde, es gibt so viele Ermessensentscheidungen – ob etwas wirklich ein Problem ist, ob es kritisch ist … Es gibt hunderte solcher Entscheidungen."

#### 3.1.2. Anzahl der Probleme

Nach der Kombination einzelner Beobachtungen zu gemeinsamen Usability Erkenntnissen wurden auf der Webseite von U-Haul 169 Probleme berichtet. Die Webseite von U-Haul ist keineswegs besonders komplex, es kann also angenommen werden, dass diese Anzahl nicht untypisch hoch ist. Der Evaluator Effect kann auch auf diese große Anzahl von Einzelproblemen zurückgeführt werden. **[Tab. 2]**

Bei der Kombination einzelner Beobachtungen zu gemeinsamen Usability Erkenntnissen hat es sich darüber hinaus als wichtig herausgestellt, die Beobachtungen sorgfältig zu formulieren. Einige Befunde beschrieben beispielsweise nur, was beobachtet wurde, erklärten aber nicht, warum ein Usability-Problem vorliegt, beispielsweise „participant is searching the faq". Andere waren ohne Kontext oder weitere Erläuterung durch den Teilnehmer gar nicht verständlich. Ein Beispiel hierfür ist folgender Befund: „to less notice about miles on the top of the page in step2 after ‚get rates' miles=city to city?". Insgesamt waren immerhin 55

| Rating | Code | Description |
|---|---|---|
| Devastating problem | AA | The problem has life-threatening or disabling consequences for users or other human beings<br>The problem could cause severe financial damages to users, the owner of the website or other persons |
| Critical problem | A | The problem causes frequent catastrophes. A catastrophe is a situation where<br>The website „wins" over the user – that is, a situation where users cannot solve a reasonable task<br>The website annoys users considerably<br>Users obtain an inappropriate solution to the task |
| Serious problem | B | Delays users in their use of the website for some minutes, but eventually allows them to continue<br>The task solution is sub-optimal and would not be accepted by users if they were informed of the „correct" solution<br>The problem causes occasional „catastrophes" |
| Minor problem | C | Causes users to hesitate for some seconds<br>The task solution obtained is sub-optimal but acceptable |
| Positive finding | P | This approach is recommendable and should be preserved |

**Tab. 3.**
Bewertungsskala aus Atlanta (CUE-9a)

| Rating | Code | Description |
|---|---|---|
| Critical problem | A | Causes frequent catastrophes. A catastrophe is a situation where the website „wins" over the test participant – that is, a situation where the test participant cannot solve a reasonable task or where the website annoys the test participant considerably. |
| Serious problem | B | Delays test participants in their use of the website for some minutes, but eventually allows them to continue. Causes occasional „catastrophes". |
| Minor problem | C | Causes test participants to hesitate for some seconds. |
| Good idea | I | A suggestion from a test participant that could lead to a significant improvement of the user experience. |
| Positive finding | P | This approach is recommendable and should be preserved. |
| Bug | X | The website works in a way that's clearly not in accordance with the design specification. This includes spelling errors, dead links, scripting errors, etc. |

**Tab. 4.**
Skala aus Chemnitz – CUE9b).

von 1.333 Beobachtungen zu unklar, um verstanden und einem Problem zugeordnet werden zu können.

### 3.1.3. Gründe für den Evaluator Effect

Als Grund für den Evaluator Effect nannten einige Teilnehmer, dass das Ziel der Evaluierung unklar gewesen sei. Es war beispielsweise unklar, wie die Verkaufsabsichten der Webseite abgewogen gegen das Interesse der Nutzer werden sollten.

Ein Beispiel dafür waren die vorselektierten Artikel im Warenkorb: Mindestens ein Teilnehmer wollte eine reine Nutzerperspektive einnehmen und war daher gegen vorselektierte Waren; andere argumentierten, dass das Ziel einer guten Evaluierung vom Kunden bestimmt werden sollte.

Ein anderer Grund für den Effekt war die Uneinigkeit darüber, in welchem Ausmaß berichtete Usability Probleme am Video des Usability Tests belegbar sein sollten. Einige Teilnehmer berichteten ein Problem nur dann, wenn die Videos dafür Belege boten, also wenn die Benutzer sich direkt beschwerten; andere berichteten auch Punkte, die sie für problematisch hielten, egal ob die Videos dafür direkte Beweise enthielten. Sie kombinierten also die Usability Evaluierung mit einer primitiven Expertenevaluierung.

Ein dritter Grund war Unsicherheit über die richtige Lösung zu Testaufgaben. Ohne Ortskenntnisse war es beispielsweise schwierig, die Antwort zu einer Aufgabe

zu kennen, bei der die nächstgelegene U-Haul-Station in der Nähe einer Adresse in der kalifornischen Stadt Fremont gefunden werden sollte. Das erschwerte es, zu bewerten wann ein Usability Problem auftrat.

### 3.2. Bewertungsskalen

Nur etwa die Hälfte der Teilnehmer in Atlanta gab an, bei ihrer Arbeit als Usability Experten eine formale Bewertung des Schweregrads von Usability Problemen vorzunehmen. Der Rest der Teilnehmer unterscheidet lediglich zwischen den wichtigsten Problemen und dem Rest. Viele Teilnehmer berichteten, dass ihnen die Bewertung ihrer Erkenntnisse schwer gefallen sei und dass die Bewertungsskalen schwer zu benutzen gewesen seien (Tabelle 3 Skala aus Atlanta – CUE9a; **[Tab. 3]**

Tabelle 4. Bewertungsskala aus Chemnitz (CUE-9b). Diese Bewertungsskala wurde erstellt weil die Ergebnisse aus Atlanta grosse Unterschiede bei den Bewertungen gleicher Probleme zeigten. Die Hypothese war dass es einfach sein würde eine bessere Bewertungsskala zu konstruieren. Diese Hypothese erwies sich als falsch. **[Tab. 4]**

Ein wichtiger Grund für diese Schwierigkeiten war dass der Bewertungsprozess verschiedene voneinander abhängige Aspekte beinhaltete, beispielsweise die Anzahl der Nutzer, die das Problem betraf, ob Nutzer in ihrer Arbeit aufgehalten wurden, ob sie frustriert wurden, ob das Problem einfach zu beheben sei, und ob ein Problem für einen realen Anwender schwerwiegende Probleme verursachen würde, auch wenn es für Testnutzer nur eine kleinere Schwierigkeit darstellte.

Tabelle 5 zeigt einige der Inkonsistenzen in Atlanta (CUE-9a). Beispielsweise wurden 28% der Probleme von einigen Teilnehmern mit AA („devastating", zu deutsch etwa: katastrophal) oder A („critical", kritisch) und von anderen mit C („Minor", zu deutsch gering) bewertet. 25% der Probleme wurden sogar von zwei oder mehr Teilnehmern als AA oder A und von zwei oder mehr anderen Teilnehmern als C eingestuft.

Tabelle 5 zeigt, dass die Änderungen an der Skala von Atlanta zu Chemnitz die Bewertungen nicht konsistenter werden ließ. In Chemnitz (CUE-9b) wurden sogar 52% der Probleme von einigen Teilnehmern als AA oder A und von anderen als C bewertet, und immerhin noch 22% wurden von zwei oder mehr Teilnehmern als AA oder A und von zwei oder mehr anderen Teilnehmern als C eingestuft **[Tab. 5]**

Ein Beispiel für ein inkonsistent bewertetes Problem war die Auswahl der richtigen LKW-Größe für den Umzug. Die Beschreibung eines LKWs auf der Webseite enthielt den Hinweis, dieser sei für ein „3-Room Apartement" geeignet. Einigen Nutzern erschien es nützlicher, die Anzahl aller Räume in der Wohnung anzugeben. Dieses Problem zeigt eine erhebliche Streuung in der Bewertung, zweimal wurde es sogar als „devastating", also AA, bewertet. **[Abb. 1]**

Die neue Skala führte also zu genauso schlechten Ergebnissen wie die erste. Darüber hinaus wurde mit der Kategorie AA ein neues Problem eingeführt. Einige Teilnehmer tendierten dazu, die höchste Bewertungsstufe für diejenigen Probleme zu wählen, die ihrer Meinung nach behoben werden sollten, unabhängig von der dramatischen Beschreibung der Kategorie („lebensbedrohend" oder „ruinierend").

| | CUE-9a | CUE-9b | Combined |
|---|---|---|---|
| Net problem findings | 134 | 93 | 169 |
| At least one AA or A, and at least one C | 23/82<br>28% | 30/58<br>52% | 42/107<br>39% |
| At least two AA or A, and at least two C | 12/48<br>25% | 7/32<br>22% | 23/65<br>35% |

**Tab. 5.**
Bewertungsskala aus Atlanta (CUE-9a)

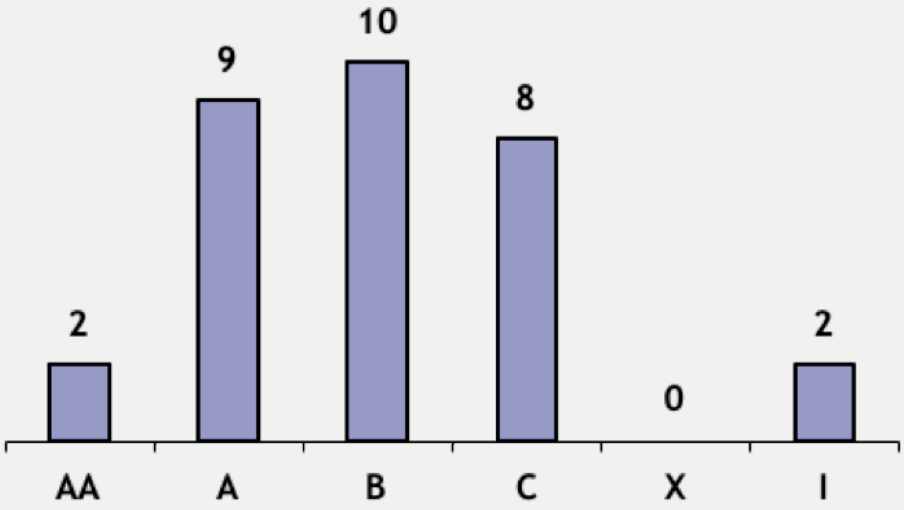


**Abb. 1.**
Bewertung des Problems „Auswahl der richtigen LKW-Größe"

## 4.
## Empfehlungen für die Praxis

Diese Studie ist für Usability Spezialisten im Arbeitsalltag in mehreren Punkten relevant. Wir geben folgende Empfehlungen für die Praxis:

**1. Die Bewertung des Schweregrad von Usability Problemen sollte in einer Gruppe stattfinden**

Eine Konsolidierung der Schweregrad-Bewertungen in der Gruppe reduziert mit einiger Wahrscheinlichkeit die Anzahl hoch bewerteter Probleme und fokussiert so die folgende Überarbeitung eines Systems.

Die gemeinsame Evaluierung in der Gruppe wurde von den Teilnehmern als eine Methode genannt, den Evaluator Effect einzugrenzen. Ein Teilnehmer fand dass „es gut ist, sich gegenseitig in einem Review-Prozess herauszufordern" und gab an, dass die Gruppenevaluierung zu konsistenteren Ergebnissen führe. In Atlanta gaben fünf Teilnehmer an, in ihrer täglichen Arbeit als Usability Experten regelmäßig zwei Evaluatoren nutzten um Usability Tests zu bewerten; die meisten anderen der Teilnehmer fanden diese Arbeitsweise zu kostspielig.

**2. Die heutigen Bewertungsskalen sind nicht verlässlich**

Die Ergebnisse der Studie zeigen, dass die heutigen Bewertungsskalen für Usability-Probleme nicht verlässlich sind. Verschiedene erfahrene Prüfer kommen zu sehr unterschiedlichen Bewertungen für die selben Probleme. Die Studie hat außerdem – unfreiwillig – gezeigt, dass es nicht einfach ist, Bewertungsskalen zu verbessern. Der Versuch, dies in CUE-9b zu tun, scheiterte. In CUE-9b haben mindestens 7 von 16 Teilnehmern ihre Befunde kritischer bewertet als nötig und die schwerwiegendste Kategorie missbraucht. Bessere Bewertungsskalen werden benötigt – natürlich müssen diese weiterhin gebrauchstauglich bleiben.

**3. Usability Probleme sollten mit besonderer Sorgfalt formuliert werden**

Bei der Überarbeitung eines Systems werden Usability Probleme oft einzeln und ohne den bei der Formulierung vorhandenen Kontext betrachtet. Die Probleme waren so teils nur noch schwer oder nicht verständlich. Die Diskussion ergab, dass Usability Experten bei der Formulierung von Usability Problemen besonders darauf achten sollten, dass aus der Formulierung des einzelnen Problems klar erkennbar ist, was das Problem ist und wie es sich auf den Nutzer auswirkt.

**4. Personen mit Domänenwissen oder Ortskenntnissen sollten zur Verfügung stehen**

Um bei der Bewertung von Benutzeraktionen Unsicherheiten zu vermeiden, sollte das Testkonzept mit Personen mit Orts- bzw. Domänenwissen abgestimmt werden und diese sollten dem Usability Spezialisten während der Auswertung als Ansprechpartner zur Verfügung stehen. Orts- und Domänenwissen kann beispielsweise notwendig werden, um korrekt zu bewerten, ob im Test eine Aufgaben richtig gelöst wurde.

**5. Usability Tests sind auch mit all ihren Schwierigkeiten sehr nützlich**

Ein Teilnehmer, der die Existenz eines Evaluator Effects akzeptierte, betonte, dass der Evaluator Effect den Wert der Usability Evaluation nicht untergraben würde: „Wir [als Evaluatoren] sind unterschiedlich. Es gibt kein endgültiges Ergebnis aber wir bieten alle eine gute Dienstleistung [für unsere Kunden]."

**6. Die Anzahl von Usability Problemen in scheinbar unkomplizierten Websiten kann hoch sein**

Unsere 35 Teilnehmer haben auf der Website von U-Haul insgesamt knapp 170 Probleme gefunden. Kein Teilnehmer hat mehr als die Hälfte dieser Probleme gefunden. Die meisten Teilnehmer haben weniger als 20% der Probleme gefunden. Die Ergebnisse der Teilnehmer sind natürlich trotzdem wertvoll, aber sie zeigen dass man niemals behaupten sollte alle oder auch nur eine Mehrzahl der Usability Probleme gefunden zu haben.

**Literatur**


1. Rolf Molich, Webseite zu den CUE-Studien, http://dialogdesign.dk/CUE.html
2. Rolf Molich, Meghan R. Ede, Klaus Kaasgaard, and Barbara Karyukin, „Comparative Usability Evaluation", Behaviour & Information Technology, vol. 23, no. 1, January/February 2004, pp. 65–74.
3. Joseph S. Dumas, Rolf Molich, and Robin Jeffries, „Describing Usability Problems – Are We Sending the Right Message", Interactions, July/August 2004, pp. 24–29
4. Molich and Joseph S. Dumas, „Comparative Usability Evaluation (CUE-4)", Behaviour & Information Technology, Vol. 27, issue 3, 2008.
5. Rolf Molich, Robin Jeffries, and Joseph S. Dumas, „Making Usability Recommendations Useful and Usable", Journal of Usability Studies, vol. 2, no. 4, August 2007.
6. Morten Hertzum, Niels Ebbe Jacobsen & Rolf Molich (2013): What You Get Is What You See: Revisiting the Evaluator Effect in Usability Tests, Accepted for publication in Behaviour & Information Technology, DOI:10.1080/0144929X.2013.783114