CUE-9

Comparative Usability Evaluation-9

Detailed Description

The Evaluator Effect

or "What You Get is What YOU See"

Workshop: Chemnitz, Germany

Sunday 11 September 2011

Call For Participation – Comparative Usability Evaluation 9

CUE-9: The Evaluator Effect Detailed Description

SUMMARY

CUE-9 will revisit the famous Evaluator Effect study originally described in *The Evaluator Effect in Usability Studies: Problem Detection and Severity Judgments* by Jacobsen, Hertzum & John from 1998.

The Evaluator Effect names the observation that usability evaluators who analyze the same usability test sessions identify substantially different sets of usability problems.

This study will investigate whether the evaluator effect still exists for experienced usability professionals and attempt to identify some of its causes.

TABLE OF CONTENTS

INTRODUCTION

The study by Jacobsen, Hertzum & John was the first research experiment done on the evaluator effect in think aloud usability studies. It demonstrated that evaluators report substantially different sets of usability problems when evaluating the same system and that they disagree about the severity of usability problems. The main explanation provided for the evaluator effect is that usability evaluation is a cognitive activity during which the evaluators must exercise judgement. This, probably, precludes complete agreement among evaluators, but the magnitude of the evaluator effect reported in the study by Jacobsen et al. is disturbing. Subsequent studies of the evaluator effect, e.g. Vermeeren et al. (2003) and Hornbæk & Frøkjær (2008), have confirmed its presence and explored additional explanations for it as well as ways of managing it. While these subsequent studies generally find a somewhat smaller evaluator effect than in the original study, it is still substantial. A better understanding of the evaluator effect is required to improve the ways of managing it.

Previous CUE studies demonstrate the evaluator effect but are not experiments. Repeating the evaluator effect study with 12 years more knowledge is important to assess its size today and the need for reducing it.

KEY QUESTIONS ADDRESSED

The key questions that CUE-9 will address are:

- What is truth, interpretation, and opinion in usability testing?
- Can the evaluator effect be replicated in 2011? How similar are the workshop participants' results?
- Precisely what is the evaluator effect? Can it be measured?
- Is the evaluator effect real, or is it wholly or partly a result of the problem matching techniques used?
- What are the causes of the evaluator effect?
- Does the moderation technique influence the reported problems? Is the evaluator effect the same in moderated and unmoderated studies?
- Are there moderation problems in the moderated usability test sessions? If so, how can these problems be prevented?
- What consensus building techniques do the groups employ at the workshop, and how successful are they?

An additional goal of the workshop is to produce and publish a set of professional usability testing videos with associated findings for teaching purposes.

APPROACH

For this study we will create two series of videos of usability test sessions of a leading edge, commercial website. Each series will consist of 5 videos. Each video covers one test session of about 30 minutes. One of the series will be from unmoderated sessions with videos made using usertesting.com. The other series will be from sessions moderated by a workshop participant. All sessions will use the same task set. The task set will be determined by the workshop participants.

Before the workshop, each participant must

- Watch one complete series of 5 videos.
- Create a list of the usability issues they would report if this was a professional study. The list must be based solely on the videos.
- Rate the severity of each identified usability problem on the list.

The anonymous lists must be submitted to the workshop organizer well ahead of the workshop. Issues can be problems or positive findings.

Workshop participants must also describe how they have analyzed the videos, in particular how they have identified and classified issues. That is, what must be happening (or not happening) in a piece of video in order for it to reveal a usability issue. Further, they must generate a list of the top moderation issues they have observed. For each issue the participant must specify the origin of the problem – that is, which session videos support the issue.

At the workshop we will analyze the participants' results and identify important similarities and differences. One of the techniques we will use is group consensus building as described in the section *Workshop Session Timeline*.

The organizer will also attempt to get access to data from the website's hotline in order to answer questions like: How do usability test results compare to the top issues reported by the hotline? Does usability testing overlook serious or critical problems? A well-managed hotline is probably the closest we can get to an authoritative source for real usability problems.

The Rashomon Effect

The Evaluator Effect is also known as the Rashomon Effect. The Rashomon effect is named after the famous 1950 Japanese crime mystery film directed by Akira Kurosawa.

The film depicts the rape of a woman and the murder of her samurai husband through the widely differing accounts of four witnesses, including the bandit/rapist, the wife, the dead man (speaking through a medium), and lastly the narrator, the one witness that seems the most objective and least biased. The stories are mutually contradictory.

The film is recommended! - even if you don't participate in the workshop.

POSITION PAPER

Workshop participants must have relevant, practical usability experience. They must have conducted at least three usability tests living up to professional standards. We will accept a limited number of workshop participants from academia and students.

A position paper is required. The position paper may be short, for example one page. It should contain the following information:

- Current affiliation
- Relevant experience in usability evaluation
- Relevant experience in consensus building for usability issues with other professionals.

Ahead of the workshop each participant must spend 5-20 hours analyzing five 30-minute videos from test sessions and report key issues. The required activities are described in more detail in the section *Pre-Workshop Participant Activities.*

The evaluator effect is about individuals. Evaluations carried out by teams will not be accepted, but we are willing to accept two or more individual workshop participants working for the same organization.

PRE-WORKSHOP PARTICIPANT ACTIVITIES

Workshop participants must evaluate videos and write a short report describing their findings ahead of the workshop.

The time plan for pre-workshop activities are:

1. 8 August 2011:
   Videos from test sessions are made available to workshop participants. Participants evaluate the videos and create a list of the usability issues they would report if this was a professional study. The list must be based solely on the videos. Participants must also rate the severity of each identified usability problem on the list.
2. 29 August 2011:
   Each workshop participant must submit
   a. An individual, anonymous report. Please use your ordinary usability evaluation report format. Workshop participants accept that their anonymous report may be made publicly available.
   b. A spreadsheet with their usability findings
   c. An addendum to the usability test report describing observed moderation issues, other information considered relevant for the study, and number of person hours that you spent on the evaluation

   Submissions may be in German or English. English is preferred.

   The report must contain at least:

   a. An executive summary
   b. List of usability findings
      - Findings must be rated. Usability problems must be rated on a scale from A (a disastrous problem) to C (a minor problem). A detailed rating scale will be provided.
      - Findings must be numbered to show the order in which they were found. Findings do not have to appear in the report in the order in which they were found.
      - For each finding the video locations supporting the finding must be provided.
3. 5 September 2011:
   All anonymous reports are made available to all workshop participants on the world wide web. The list of workshop participants is published.

   Just before the workshop, participants should spend at least 3 hours familiarizing themselves with the findings reported by other workshop participants.

WORKSHOP SESSION TIMELINE

Timeline	Topic or Event
09.00 to 09.30	Introduction to workshop. Brief presentation of each participant.
09.30 to 10.30	The individual evaluator effect. Four to six participants meet in a group for consensus building. Participants will discuss a subset of the findings they have reported and attempt to build consensus on findings and their ratings.
10.30 to 11.00	Break.
11.00 to 12.30	The group evaluator effect. Plenum discussion of selected usability findings and ratings. Why are some findings reported by some teams and not by others? Why are identical findings rated differently by teams?
12.30 to 14.00	Lunch break.
14.00 to 15.00	Discussion. Lessons learned - Criteria used by evaluators for detecting problems - Problem matching techniques - What we can do to minimize the evaluator effect
15.00 to 15.30	Conclusion. Further work.

ABOUT CUE

CUE-9 is the ninth in a series of Comparative Usability Evaluation (CUE) studies. Previous studies were conducted from 1998 to 2009. The essential characteristic of a CUE study is that a number of commercial and academic organizatons involved in usability work agree to evaluate the same product or service and share their evaluation results at a workshop. Previous CUE-studies have focused mainly on qualitative usability evaluation methods such as think-aloud testing, expert reviews, and heuristic inspections. CUE-8 focused on usability measurement.

For an overview of the eight CUE-studies and their results see www.dialogdesign.dk/CUE.html.

CUE-9 differs from previous CUE-studies in important ways. All evaluations are based on the same task set. Evaluations are based on pre-recorded videos of test sessions. The design of the study eliminates task selection and to some extent different samples of test participants as causes for differences in reported issues.

ORGANIZER

- Rolf Molich, DialogDesign (Denmark), molich@dialogdesign.dk