
You Say 'Disaster', I Say 'No Problem': Unusable Problem Rating Scales

Rolf Molich

DialogDesign
Skovkrogen 3
DK-3660 Stenlose, Denmark
molich@DialogDesign.dk

Jennifer (Jen) McGinn

Oracle Corporation
10 Van De Graaff Drive Burlington,
MA 01803, USA
jen.mcgin@oracle.com

Nigel Bevan

Professional Usability Services
12 King Edwards Gardens
London, W3 9RG, UK
nigel@nigelbevan.com

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CHI'13, April 27 – May 2, 2013, Paris, France.

Copyright 2012 ACM 978-1-XXXX-XXXX-X/XX/XX...\$10.00.

Abstract

This paper documents a consistent and considerable evaluator effect in the ratings of usability problem severity carried out by experienced usability professionals. The CUE-9 study, conducted in 2011, showed a reasonable level of rater agreement in problem identification, but the severity assigned to problems varied wildly from "not a problem" to "disaster". This paper documents the variations we observed and the need for a better rating scale and better training. The paper calls for more caution when using extreme ratings. Our results also show that simple attempts to fix the traditional rating scales may not work.

Author Keywords

Usability; problem rating; rating scale; severity; problem severity; rater agreement; usability problem; usability evaluation; evaluator effect.

ACM Classification Keywords

H5.2.e. Information Interfaces and Representation (HCI), User Interfaces, Evaluation/Methodology.

General Terms

Human Factors; Design; Measurement.

INTRODUCTION

Over the last 20+ years, numerous rating scales have been suggested to rate usability problem severity and the resulting priority for fixing those problems.

A survey by Hertzum [4] of five previous studies found a lack of consistency in usability problem ratings that is similar to what this paper reports, with weak correlations among the severity ratings of experienced evaluators.

Since 1998, the Comparative Usability Evaluation studies [2] have examined how results of usability evaluations differ as a result of method selection, task design, number of participants, and individual practitioner differences. The most recent of these studies, CUE-9 [5], attempted to control for more variables than any of the previous CUE studies by having professional evaluators observe the same videos and then rate the usability problems using the same scale. Evaluators thus used the same participants completing the same tasks. The goals of CUE-9 were to determine if the evaluator effect [6] was still a problem with regard to reproducibility, and if so, what factors contributed to that effect.

THE CUE-9 STUDY

In the CUE-9 study, 35 professional evaluators were given 5 videos of participants working through 5 typical tasks on the U-Haul website, U-Haul.com. U-Haul's main business is to rent moving trucks and trailers. Each video had a length of approximately 30 minutes. Evaluators were asked to evaluate 5 videos and describe their usability findings in a report that they would submit to the presumed client (U-Haul). Two sets of 5 videos each were involved - set 1 was moderated

while set 2 was unmoderated. Both sets used the same test script. No limits were imposed on how long evaluators could take to conduct and report their analysis, and no specific report format was mandated. Evaluators were provided a rating scale to use when categorizing their usability problems. Further details about CUE-9 can be found in [2] on the CUE-9 page.

CUE-9 part	Conducted	Number of Evaluators	Video set	Rating scale
a	April 2011	9	1	Table 2
a	April 2011	10	2	Table 2
b	August 2011	16	2	Table 3

Table 1. Variations of videos and rating scales in CUE-9. The first CUE-9 study, CUE-9a, was conducted in April 2011. The interest in CUE-9a was considerable, so we decided to run a similar study, CUE-9b, which took place in August 2011.

RATING SCALES

The rating scale used in CUE-9a was based on similar scales used by Barnum [1], Dumas [3] and Nielsen [7].

Our intent was to create a rating scale that

- Was usable. We wanted a short and simple scale, so we limited ourselves to three problem levels: critical, serious and minor with short explanations.
- Provided for positive ratings
- Focused strictly on usability goals. Rating scales may confound usability goals by including phrases such as "imperative to fix this" or "must fix". While the usability specialist can definitely argue that a problem is critical for the user, the final decision rests with management who takes into consideration time to fix, cost and business impact.

The CUE-9a study used the following rating scale.

Rating		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.

Table 2. Usability rating scale used in CUE-9a

A preliminary analysis of the reports from CUE-9a showed some shortcomings of the problem ratings:

- Ratings varied wildly. For example, 12 problems were rated A by 2 or more evaluators and C by 2 or more other evaluators. See Table 4.
- The usability impact of problems rated X and I was not available. Even if a finding is a bug or an idea voiced by a user, we need to know its usability impact. Also, while X and I are valid classifications, they do not fit into a rating scale from positive to negative.

- Some important problems are worse than “Critical problem”. We need to be able to distinguish between highly inconvenient showstoppers and problems that have life-threatening or disabling consequences for users or other human beings.

Rating		Description
Devastating problem	AA	<ul style="list-style-type: none"> • The problem has life-threatening or disabling consequences for users or other human beings • The problem could cause severe financial damages to users, the owner of the website or other persons
Critical problem	A	<p>The problem causes frequent catastrophes. A catastrophe is a situation where</p> <ul style="list-style-type: none"> • The website “wins” over the user – that is, a situation where users cannot solve a reasonable task • The website annoys users considerably • Users obtain an inappropriate solution to the task
Serious problem	B	<ul style="list-style-type: none"> • Delays users in their use of the website for some minutes, but eventually allows them to continue • The task solution is sub-optimal and would not be accepted by users if they were informed of the “correct” solution • The problem causes occasional “catastrophes”
Minor problem	C	<ul style="list-style-type: none"> • Causes users to hesitate for some seconds • The task solution obtained is sub-optimal but acceptable
Positive finding	P	This approach is recommendable and should be preserved

Table 3. Modified usability rating scale used in CUE-9b

Disaster vs. No Problem

Here are some of the key arguments voiced in support of the extreme ratings of "The Problem that Wasn't":

Disaster: "The videos show that users do not notice or understand the differences between the insurances offered. If I hit someone while driving a U-Haul truck and it is my fault, disaster could strike. If I have unknowingly chosen the insurance with inadequate coverage, I could be personally liable for hundreds of thousands of dollars. This would ruin me."

No problem: "While I think it would be *desirable* for the corporation to try to educate its customers about how to wisely choose insurance, I don't think it's their responsibility, nor would their efforts necessarily be effective. They have a reasonable expectation that anyone renting a vehicle knows that there is a chance of an accident, and that that is why they have the option of purchasing insurance."

We considered these shortcomings sufficiently serious to warrant a revision of the rating scale. The 16 participants in the CUE-9b study used the modified rating scale shown in Table 3. None of the CUE-9b evaluators expressed a need for the "Bug" or "Idea" ratings during or after the study.

RESULTS

The 35 evaluators reported a total of 1,332 findings. The first author and Morten Hertzum independently combined similar CUE-9a findings into groups. The CUE-9b results were subsequently classified using the same groupings. The process is described in [5].

For the combined CUE-9 studies, this categorization resulted in 222 groups of findings of which 53 were positive and 169 were usability problems. Of the 169 usability problems, 14 were reported by more than half of the participating evaluators.

CUE-9 Results Summary	CUE-9a	CUE-9b	Combined
Number of evaluators	19	16	35
Total number of submitted findings	859	473	1,332
Net total number of findings after combining similar findings	182	109	222
Net positive findings	48	16	53
Net problem findings	134	93	169
Problems reported by single evaluators only	52/134 39%	35/93 38%	62/169 37%
Problems reported by 4 or more evaluators	48/134 36%	32/93 34%	65/169 38%
At least two AA or A, and at least two C	12/48 25%	7/32 22%	23/65 35%

Table 4. Key rating results from CUE-9. The *Combined* column shows the number of unique problems identified in 9a and 9b.

The Problem that Wasn't

One U-Haul usability problem illustrated how evaluators used different criteria when making ratings: During the rental process, the language of the user interface encouraged the user to select a cheap van insurance option, but did not inform the user that this cheap insurance option provided vastly inadequate coverage if the user was found at fault in an accident.

Initially, only one participant reported this problem. After the main part of the study had been completed, we asked all evaluators to rate this particular problem (see the sidebar for some explanations). The ratings varied considerably:

- 5 evaluators rated the problem AA, including 2 evaluators who had not rated any other problem AA
- 4 evaluators rated the problem A
- 5 evaluators rated the problem B or C
- 6 evaluators said that they did not consider the problem a usability problem
- 15 evaluators didn't respond or returned unclear answers.

Consistently Inconsistent Ratings

Table 4 shows the inconsistencies in the problem ratings in CUE-9. We have combined the findings from CUE-9a and CUE-9b in Table 4 to show that the changes we made to the rating scale from CUE-9a to CUE-9b did not make ratings more consistent.

In CUE-9a 25% of problems were rated A by 2 or more evaluators and C by 2 or more other evaluators. This figure was more or less the same (22%) in CUE-9b.

The new rating scale yielded as much divergence in severity as the previous one.

Inappropriate AA Ratings

The AA rating was added to the rating scale to make the scale more widely applicable and to test the hypothesis that some evaluators exaggerate ratings.

Only problems that have life-threatening, disabling or major financial consequences to the end user should be rated AA. Seven of the 16 CUE-9b evaluators (44%) applied the AA rating. They used it 23 times on 16 problems. Each of the 7 evaluators who used the AA rating reported 1 to 6 findings that they classified as having a severity of AA. No problem was rated AA by more than 2 evaluators so the evaluators did not agree on the AA-ratings.

What caused the evaluator effect in CUE-9?

We have identified a number of possible reasons for the inconsistent ratings in CUE-9, each of which are related to the rating scales provided.

a) Inadequate descriptions of the scale labels

The scale does not describe how certain important problem types should be rated. For example, how should an evaluator rate a problem where omitting or hiding information might hurt sales, or a problem that prevents one user from completing a key task while 4 other users only have minor problems?

The scale items are hard to understand because they do not include specific examples.

b) Multiple aspects of usability problems were combined
A single letter grade includes the frequency, impact and persistence of a problem. Different evaluators may have weighted these criteria differently when making their judgments. Although using separate scales for each of these criteria might be more reliable, combining them to produce an overall rating would still need a basis for weighting or prioritizing each scale.

c) Taking account of consequences

The scale description focused on whether users could achieve their immediate interaction task, but as the insurance problem illustrates, the rating of some evaluators was influenced by the impact of the *potential* consequences of a usability problem. Another example that illustrates this point would be the potential consequences of a user renting a van that is too small.

d) Disregarding the descriptions

We speculate that the inappropriate AA Ratings were a result of the evaluators not reading or disregarding the definition of the AA rating and wanting to indicate that a problem belonged to the most serious category on U-Haul's website ("must fix").

Conclusion

Usability problem ratings using a traditional rating scale are unreliable. Different experienced professionals arrive at ratings that differ wildly for the same problems.

Trying to make a better scale failed. At least, we failed in our first attempt. Our study also - involuntarily - demonstrates that creating a better rating scale is not simple.

Some of the UX professionals exaggerated their ratings. At least 7 out of 16 evaluators in CUE-9b exaggerated ratings. This could influence trust in ratings and in usability studies in general.

Better usability rating scales and perhaps better training is needed - of course, without sacrificing the usability of the scales.

Future Work

- Identify and measure basic criteria

One reason for the differences may be that the criteria contributing to a rating (such as the severity or frequency of a problem) are given different importance by different evaluators. We plan to identify the independent criteria and measure them using separate scales to find out whether these are rated more consistently. We will then use what we have learned to design one or more new scales that better explain the relative importance of the different criteria. We are planning a comparative evaluation of the new and old scales.

- Test the usability of the new scale

The length of the description of the scale and the comprehensibility of the description affect the usability of the rating scale. To be usable, the scale may also require supplementary notes and better examples to elaborate the rating descriptions.

- Focus on the bigger process: Train prospective raters and experiment with the rating process

Training combined with some kind of test may be required to ensure that people who aspire to rate problems fully understand the scale.

Training examples may be needed to enable prospective raters to check their understanding of the

scale.

We may get more consistent results by asking testers to rate the severity of problems in private and then discuss the ratings in a group.

- Check: Have we built a better mousetrap?

At the end of the day we must prove convincingly that the new scale is indeed superior to existing scales.

ACKNOWLEDGMENT

We sincerely thank the 35 evaluators for investing considerable time in CUE-9. Their names and affiliations can be found in [2]. We also thank Morten Hertzum for independently classifying and checking the 1,332 findings in cooperation with the first author.

REFERENCES

1. Barnum, C.M. *Usability Testing Essentials*. Morgan Kaufmann, Boston, MA, USA, 2011, 263–265.
2. CUE - Comparative Usability Evaluation Overview <http://www.dialogdesign.dk/CUE.html>
3. Dumas, J. & Redish, G. *a practical guide to usability testing*. Ablex, Norwood, NJ, USA, 1999, 324–326.
4. Hertzum, M. Problem prioritization in usability evaluation: From severity assessments toward impact on design. *International Journal of Human-Computer Interaction* 21, 2 (2006), 125–146.
5. Hertzum, M., Molich, R. & Jacobsen, N.E. What You Get Is What *You* See: Revisiting the Evaluator Effect in Usability Tests. Submitted for publication.
6. Jacobsen, N.E., Hertzum, M., & John B.E., The evaluator effect in usability studies: Problem detection and severity judgments. *Proceedings of the Human Factors and Ergonomics Society*, 1998, 1336 – 1340.
7. Nielsen, J. *Usability Engineering*. Morgan Kaufmann, Boston, MA, USA, 1993, 103–104.