Cooperative Usability Testing: Complementing Usability Tests with User-Supported Interpretation Sessions

Erik Frøkjær & Kasper Hornbæk
Datalogisk Institut, Københavns Universitet
Universitetsparken 1, DK-2100 Copenhagen, Denmark
{erikf, kash}@diku.dk

ABSTRACT

Recent criticism of think-aloud testing (TA) discusses discrepancies between theory and practice, the artificiality of the test situation, and inconsistencies in the evaluators' interpretation of the process. Rather than enforcing a more strict TA procedure, we describe Cooperative Usability Testing (CUT), where test users and evaluators join expertise to understand the usability problems of the application evaluated. CUT consists of two sessions. In the interaction session, the test user tries out the application to uncover potential usability problems while the evaluators mainly observe, e.g. as in TA or contextual inquiry. In the interpretation session, evaluators and test users discuss what they consider the most important usability problems, supported by a video of the interaction session. In an exploratory study comparing CUT to TA, seven evaluators find that interpretation sessions contribute important usability information compared to TA. Also test users found participation in the interpretation session interesting.

Author Keywords

Usability testing, think aloud, metaphors of human thinking, contextual inquiry

ACM Classification Keywords

H.5.2 [Information Interfaces and Presentation (e.g., HCI)]: User Interfaces—Evaluation/Methodology

INTRODUCTION

Think-aloud testing (TA) has been used for more than 20 years to identify usability problems with an interface [4]. Usually, think-aloud tests let the user solve typical tasks with the interface while continuously verbalizing his or hers thoughts. The user is then debriefed and one or more evaluators interpret the users' think aloud to predict usability problems with the interface. Because TA is simple, and gives an immediate, powerful experience of users' difficulties, it is the most popular technique for usability evaluation [12].

The validity and practical use of think-aloud as a usability evaluation technique, however, has recently been criticized. Boren and Ramey [2], for example, have pointed out that while Ericsson and Simon's Verbal Protocol Analysis [5] is often cited as the theory underlying TA, practical TA tests are much less rigorous and often in contradiction with the recommendations of Ericsson and Simon. Another line of critique has pointed out that the test situation prescribed by TA is artificial and restricts constructive dialogue with the user [3]. Finally, work on the evaluator effect suggests that observers of TA find markedly different problems [6]. While usability testing in a sense appears objective, this work suggests that interpretation ultimately shapes what gets reported as usability problems. Altogether, these critiques suggest that rethinking of TA is pertinent.

We present a technique, called Cooperative Usability Testing (CUT), for complementing usability testing with user-supported interpretation sessions. CUT relaxes the requirements of think-aloud testing, while enabling the user to influence the interpretation of the test. Thus, CUT aims to help identify usability problems with better validity than those found with TA, and to improve users' and evaluators' experience of participating in a usability test. Our overall intention has been to organize usability testing as a cooperative learning process.

COOPERATIVE USABILITY TESTING (CUT)

In CUT, the usability test of a system is done in cooperation between test users and evaluators. The test users' expertise concerns the work domain, including experience with alternative systems. The evaluators' expertise concern HCI and usability, often supplemented by experience in systems development. The chief idea of CUT is to bring together users and evaluators in a constructive dialogue aimed at uncovering usability problems. This happens through (a) an *interaction session* directed by a test user, who performs relevant tasks with the system to uncover usability problems, and (b) a cooperative *interpretation session*, directed by evaluators, based upon a video of the interaction session. These two sessions are explained in the following subsections.

Interaction session (IAS)

The interaction session (IAS) can be conducted for example as a think aloud test [9] or as a contextual inquiry [1]. It is vital that the test user has the initiative in the IAS. The test user can ask the evaluator for assistance, for example if

CHI 2005, April 2–7, 2005, Portland, Oregon, USA.
ACM 1-59593-002-7/05/0004.

serious problems are experienced when trying to do the tasks planned or if the application crashes. When the application evaluated only exists as a paper prototype or electronic mock-up, the evaluator also needs to play an active part of the session.

One concrete way to do the IAS is to use two evaluators, a guide and a logger, together with one test user. The procedure of the evaluation and the roles of the evaluators are briefly described to the test user at the beginning of the session. Evaluators should make it clear to the test user that he/she is the primary active person. The entire IAS should last at most 45 minutes. It is important that the entire session is video-taped in such a way that the sound is clear and the application visible.

Interpretation session (IPS)

The interpretation session (IPS) is conducted in cooperation between the test user and the evaluators. It aims to identify and understand the most important usability problems brought up in the previous IAS.

Professional usability evaluators usually have a standard procedure for identification and understanding usability problems within the domains and interaction styles they usually work with. It would be natural to take guidance from that procedure. Sometimes the evaluators are trainees or system developers with little experience in uncovering usability problems. In these cases the IPS might be supported by a solid usability inspection technique. We suggest as one possibility metaphors of human thinking, MOT. MOT has been shown useful to describe a broad range of HCI phenomena, and has performed well in comparison to heuristic evaluation [7] and cognitive walkthrough [8].

The IPS should not aim at producing complete descriptions of usability problems, but only establish a clear understanding of the most important and complicated usability problems. The details of problem descriptions can be filled in during the evaluators' reporting of problems.

One concrete way to conduct the IPS is to do it after a short break following the IAS. The IPS should last maximally 45 minutes. Compared to the IAS, it makes sense for the guide and logger to switch roles, so that the evaluator who played the role of logger in the IAS can use the notes taken during that session. The person who is acting as logger in the IPS should facilitate navigation in the video and take notes about the discussions between the guide and the test user. The logger should also prevent the guide from going into too many details about minor issues of the IAS.

While going through the video the guide points out the sequences expected to shed light on important usability problems. The pointing out are done in cooperation with the test user. Both the guide and the logger should pay careful attention to whether the test user actively shows agreement in choosing a certain sequence for scrutiny. This will not always be explicitly stated. If the guide feels that the test user is passive and does not appear to find a problem important, then he should—without much discussion or explanation—proceed to the next segment of interest. The limited time with the test user must be used as effectively as possible to cooperate on interpreting the sequences of the video with the important usability problems.

Related work

CUT differs from existing descriptions of TA, e.g. [9], by including an interpretation session and by relaxing the requirement that the evaluator should affect the test as little as possible. Compared to retrospective testing, e.g. [10, 11], the evaluator tests hypotheses in the IPS, some of which could only be formed because the test user was thinking aloud during the interaction session. CUT seems to be unique in the approach of combining systematic interaction and interpretation sessions.

EXPLORATORY STUDY

We performed an exploratory study of CUT by having seven evaluators use the two evaluation techniques CUT and TA. The aim is to collect initial data on how the interpretation session works and is experienced by evaluators and test users. Note that in this study, the IAS of CUT is performed using think aloud—the main difference between the two evaluation techniques is whether or not an interpretation session is present, see Figure 1.

The evaluators performed their evaluations in pairs, except one who conducted the evaluations by himself. Evaluators used both CUT and TA; the order in which the techniques were used was counter-balanced. The evaluators evaluated two parts of http://www.dr.dk, the web site of the national Danish broadcasting company (DR). They were given ten tasks for each part, developed in cooperation with DR.

TA was described to evaluators by [9]. Further, we asked that TA lasted maximally 45 minutes. CUT was described to evaluators by a document similar in content to the previous section, but containing more practical recommendations; evaluators were asked to follow a think aloud procedure [9] in their IAS. Evaluators were instructed to use a maximum of 45 minutes for the IAS. After this session, they could take a break less than 15 minutes. The IPS should last at most 45 minutes. During both CUT and

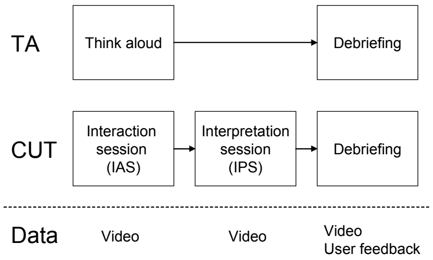

Figure 1: The sessions in the use of TA (top row) and CUT (middle row). The bottom row shows what data are collected from the various sessions of this study.

TA evaluations, one evaluator acted as the guide, talking to the test user; the other logged the usability problems discovered. A total of eight test users participated in the CUT and TA evaluations; their sex and professional backgrounds were mixed.

After each evaluation, evaluators were instructed to hand in: (a) a list of usability problems, including an assessment of severity; and (b) test user feedback about participation in the evaluation, including whether it felt natural to think aloud and if the interpretation session focused on relevant issues. In addition to this feedback, we videotaped those parts of the evaluation where the test user was present. We refrain from investigating the usability problems identified because the data is too limited to make statistical analysis meaningful.

RESULTS

Below we summarize the comments of the test users and our analysis of the videos of the IPS.

Test users' perception of the evaluation

All test users made one or more comments on negative aspects of thinking aloud. Comments made by more than one participant include (a) that it is hard to think aloud while reading and scanning text, (b) that it is hard to keep thinking aloud when tasks are challenging, and (c) that what can be said out loud felt as only a fraction of what the test user was considering. Also, test users described the thinking aloud as "asocial monologue" and "stressful".

The test users participating in IPS made mostly positive comments. Two of them considered it nice to be able to explain their actions; other two commented that it was interesting to take part in the interpretation. Also one test user expressed that it was interesting to reflect upon the IAS. In contrast to the majority of test users, one person did not feel that it was really relevant for him to participate in the interpretation session.

Videos of the interpretation session

The most important material in this exploratory study of CUT was the videos of the IPS. The authors looked through the videos, noting important episodes in the interpretation with a description and a time-stamp. To characterize how the interpretation session works, we below focus on (1) whether the video of IAS was presented in full length, or only selected sequences were presented; and (2) whether the guide in relation to the test user took a mainly active role or a mainly reactive role. Other important observations are summarized in Figure 2.

Two groups (G3 and G4) chose to let the full video of the IAS session be presented as basis of their IPS. Sometimes, especially in the beginning of the IPS, the video was stopped to give time for discussing certain usability problems. Winding the video forward to next interesting sequence was used very rarely. Two other groups (G1 and G2) chose to focus only on sequences that to the evaluators had appeared to contain the most important usability problems, and these groups made extensively use of wind forward. G1 started their IPS with a brief introduction of the selected 3-4 situations which they proposed to focus on.

Strengths	Opportunities
Test users and evaluators experience their dialogue as natural and meaningful. Interpretation sessions give test users a context-based feedback and debriefing which seems especially satisfying for very serious usability problems and test tasks that remained unsolved. The interpretation sessions let test users comment across their experiences of the IAS. The evaluators' review of the IAS video and the direct discussions with the test user offer effective learning processes for the evaluators to improve their understanding of users and their performance of TA and CUT evaluations. The interpretation session facilitates discussing of ideas for solutions addressing the usability problems raised.	Improved facilities for winding and retrieval of wanted video sequences might support a selective video presenting approach. Combine interaction with the tested system and viewing the video. Reviewing all of the video during the IPS works fine. Discussions about important usability problems can continue while the less interesting parts of the IAS session are being presented in the background. Techniques for taking notes during CUT needs to be developed and trained to make notes more useful in supporting the coherence of the evaluation process.
Weaknesses	**Threats**
The interpretation sessions sometimes lead to discussions focused upon the interaction processes leaving less attention to the interpretation and understanding of usability problems. Winding of the video to find the wanted sequences is rather time-consuming and will often direct the attention of all participants to this secondary problem. A systematic interpretation approach is not yet developed. The MOT usability inspection technique showed inadequate for the highly interactive and adaptive communication processes of the IPS.	The evaluators are at risk to introduce new and maybe problematic interpretations. This risk might in some situations be more threatening than in TA because interpretations are expressed and discussed very rapidly without careful analysis. The guide's demanding work conditions during the IPS makes it hard to utilize the notes taken during the IAS. There is a risk that discussions during the IPS are being too general and partly detached from the IAS.

Figure 2: Strengths, weaknesses, opportunities and threats introduced by the interpretation sessions of CUT.

G2 made their selections less explicit, and as the IPS proceeded this group gradually had their video running without breaks, similar to G3 and G4.

Our impressions from reviewing the IPS videos are that replay of the IAS in full length gives a more easygoing, yet very focused session compared to presenting only the important usability problems. The pauses with winding for retrieval of the next sequence for discussion seem to invite the guide to motivate and set the scene for the interpretation in ways that sometimes lead into rather general and hypothetical discussions of the issue at hand. Winding of the video, although handled by a second evaluator, was with our equipment both a time-consuming and, more importantly, a disturbing activity which attracted the attention and participation also of the test user and the guide. The rather long periods without really interesting interaction taking place, which we had foreseen to be a problem when creating the CUT technique, seem to be used very naturally for interesting discussions and sharing of experiences between the guide and the test user. The second evaluator now and then falls in with questions and possible new interpretations. We were surprised to see how all participants in this study, that is both the evaluators and the test users, demonstrated that it was quite easy to concentrate and exchange interpretations while the IAS video continued running.

The degree of active participation by the guide of the IPS was the other very influential aspect of the IPS process. The groups G1 and G3 seemed to have more active guides who to a larger extent stated their comments to and interpretations of what they thought to have observed during the IAS. In the two other groups, the guides used a more cautious and reactive attitude in letting the test user be the one taking more initiative and control of the IPS.

Our impressions from the video are that the active guide strategy seems to promote a more thorough and challenging dialogue between the test user and the evaluators. In a number of cases learning processes involving all participants were demonstrated. Something similar might have taken place more implicitly during the reactive guide strategy, but this can not be documented by our study. The active guide strategy seems to invite and involve the test user in a meaningful dialogue about what happened during the IAS. Test users and evaluators gain clearly from sharing the participation in the IAS just a few minutes earlier.

CONCLUSION

Cooperative usability testing (CUT) combines interaction and interpretation sessions where test users and evaluators join expertise to understand the usability problems of the application evaluated. This seems to circumvent problems associated with traditional think-aloud testing on the one hand, and retrospective usability testing on the other.

Test users of CUT reported that they liked being able to reflect and comment upon their interactions, compared to think-aloud studies even with extensive debriefings. Evaluators likewise thought that the interpretation session was valuable to clarify and understand usability problems. Yet, the interpretation session was challenging for evaluators. They found it hard to utilize their notes about the users' interaction. There was also a risk that discussions became too general and detached from what happened during the interaction session.

Based on this initial study of CUT we advise evaluators to guide the interpretation session actively and be cautious using a selective style of presenting the video. Further advice might be taken from the summary in Figure 2.

Further work is needed to develop better support for the interaction and interpretation sessions of CUT, for example techniques for cooperative interpretation of usability problems. Such support must respect the highly interactive communication processes characterizing CUT.

ACKNOWLEDGEMENTS

The study was done in collaboration with our students Lars Christiansen, Jonas Fonseca, Kasper Hansen, Claus Jensen, Kaare Kaaber, Christina Nielsen, and Solveig Søndergaard; and with Sofie Scheutz from DR.

REFERENCES

1. Beyer, H. & Holtzblatt, K. *Contextual Design*, Morgan Kaufman Publishers, 1998.
2. Boren, M. T. & Ramey, J. Thinking Aloud: Reconciling Theory and Practice, *IEEE Transactions on Professional Communication*, *43*, 3 (2000), 261-277.
3. Buur, J. & Bagger, K. Replacing Usability Testing With User Dialogue, *Communications of the ACM*, *42*, 5 (1999), 63-66.
4. Dumas J. S., User-Based Evaluations, in Jacko, J. A. & Sears, A. *The Human-Computer Interaction Handbook,* Lawrence Erlbaum Associates, 2003, 1093-1117.
5. Ericsson, K. A. & Simon, H. *Protocol Analysis: Verbal Reports As Data, Revised Edition*, MIT Press, Cambridge, MA, 1993.
6. Hertzum, M. & Jacobsen, N. E. The Evaluator Effect: A Chilling Fact About Usability Evaluation Methods, *International Journal of Human-Computer Interaction*, *13* (2001), 421-443.
7. Hornbæk, K. & Frøkjær, E. Usability Inspection by Metaphors of Human Thinking Compared to Heuristic Evaluation, *International Journal of Human-Computer Interaction*, *17*, 3 (2004), 357-374.
8. Hornbæk, K. & Frøkjær, E. Two Psychology-Based Usability Inspection Techniques Studied in a Diary Experiment, *Proc. Nordichi 2004*, ACM Press (2004)
9. Molich, Rolf, User testing, Discount user testing, 2003, www.dialogdesign.dk.
10. Nielsen, J. *Usability Engineering*, Academic Press, San Diego CA, 1993.
11. Nielsen, J., Christiansen, N., Clemmensen, T., & Yssing, C. Mindtape-a Technique in Verbal Protocol Analysis, *HCI International*, Lawrence Earlbaum Associates (2003), 188-192.
12. Vredenburg, K., Mao, J.-Y., Smith, P. W., & Carey, T. A Survey of User-Centered Design Practice, *Proc. CHI 2002*, ACM Press (2002), 472-478.