



Understanding Usability Work as a Human Activity

Nørgaard, Mie

Publication date:
2008

Document version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Nørgaard, M. (2008). *Understanding Usability Work as a Human Activity*. København: Department of Computer Science, University of Copenhagen.

human perspective,
usability work from hu.
of evaluation results. This
usability as a human activity inv.
collaborating to improve usability. In the
ish translation see (Naur, 1992), and lat
understanding work on computer systems. The
ork on usability concern methods, procedures,
course, work on how to describe and present resu.
e receivers. Yet, work that aim to isolate import
methods (UEMs), seem often to view usability deta
. To improve the downstream utility of usability
o understand usability in terms of activities such a
collaborating with other stakeholders to improve usability
om [understanding usability work as a human activity](#) (Nau
be understood as the development and use of certain ev
onality, individual professional goals, learning, and collabo
on usability views usability work and results from a di
understood the results of usability evaluations as a presentation of
for example (Redish, Bias, Bailey, Molich, Dumas, & Spool, 2002;
suasiveness has been mentioned as a key factor for usabil, 's
y terms such as relevance, salience, reliability and quantity (L
ised by Nørgaard and Høegh (2008 in terms of argumentation th
y results concerns for example the value of redesign proposals (H
ts (Hertzum, 2006). Furniss et al. (2007a) relate themes from Re
argue that usability is an activity that needs to be adjustable and fle
anging contextual factors. The same view has been touched upon by Nørg
t usability experts feel the need to adjust their usability testing to chang
y work to the organisation in which it takes place, and suggests that usabil.
od together with other activities and contextual factors, and not as a star
t of context. The work discussed above, which is mainly relate
al factors, clearly has a human perspective. For example,
inspire developers, shows concern for the humans who receive
usability as a human activity that is dependent on how huma
ow the role of the human is discussed in usability literature. r
ol, 2002) has argued that the relationship between develop
factor for usability's success, more important than, for example
s. Others have made similar observations on the impor
p between usability expert and customer, users and stakeh
ndford, & Curzon, 2007b; Wixon & Wilson, 1997). In fact,
facilitate collaboration between stakeholders to usability
lines I suggest that researchers and practitioners should u
a matter of methods and procedures. Such an understandi
collaboration and learning between stakeholders. It max
ls but as individuals who work together in -
dition, with its strong focus on collp
an activity (Bødker & Buur, 2002)
s, where much has the form of
ample the evaluation of protc
activity will help researchers
the sketching phase of dr
tion. In my opinion we
o better understand
bility experts, and
individuals as well. c
somewhat inward focus
e of implementing usabi.
olders learn about and ur
r focus will move usability
tifies, or how many of the i
is needed to bring qualitativ
professionals conduct usability wor
attention, see for example (Cajano
004; Uldall-Espersen & Frøkjær, 200
rk is organised and conducted in differ
s and goals for the usability of products
reflect on usability work in an organisati
n improvements came to be, would also be
se questions (and many more) researchers ne
usability is about UEMs, redesigns, problem se
I argue that usability work is mainly about huma
o improving usability is not mainly about getting
ems in more detail. Improving usability is about un
process, where both professional and personal relatio
e complex nature of usability work, and how we mig.

Under- standing Usability Work as a Human Activity

MIE NØRGAARD

PHD THESIS

DEPARTMENT OF COMPUTING

FACULTY OF SCIENCE

UNIVERSITY OF COPENHAGEN

SEPTEMBER 2008

Content

Abstract	2	Papers	15
Dansk Sammenfatning	3	What Do Usability Evaluators Do in Practice? An Explorative Study of Think-Aloud Testing	18
Usabilityarbejde forstået som en menneskelig aktivitet	3		
Usabilitypraksis og metoder	4	Usability Work: A Human Activity	31
Overbevisningskraft og det at blive overbevist	4	User Testing in the Combat Zone	35
Preface	6	Working Together to Improve Usability: Challenges and Best Practices	41
Introduction	7	Exploring the Value of Usability Feedback Formats	57
Understanding usability work as a human activity	7	Evaluating Usability - Using Models of Argumentation to Improve Persuasiveness of Usability Feedback	73
Practice and methods	11	Can Eyetracking Boost Usability Evaluation of Computer Games?	87
Persuasiveness and the process of being persuaded	11	Organizational Challenges to User Research in the Video Game Industry: Overview and Advice	93
Future work	12		
References	12		

Abstract

Three core themes are explored in eight papers: Usability work as a human activity, usability practice and methods, and persuasiveness of evaluation results and feedback. We explore how usability work is much more than methods and work procedures, and argue that maturing our understanding of usability work to include a human perspective, is crucial to downstream utility—how usability work impacts the on-going development process. Our work shows that cross-professional collaboration is subject to challenges that arise from stakeholders having conflicting priorities, procedures and personalities. Such challenges include evaluation results lacking relevance, poor timing of evaluation results, little respect for other disciplines, and difficulties sharing important information about a design. The studies of practical usability work suggest that user researchers working with computer games and task oriented

systems struggle with making methods meet practical realities and demands, and that the concept of usability in games is not satisfactorily covered by for example the ISO 9241-11. With this in mind we call for future work that broadens the concept of usability to include concepts more relevant to games—such as fun and aesthetics—and explores evaluation methods that reflect such aspects. Our focus on persuasiveness suggests that persuasiveness is not an attribute of certain feedback formats. We have conducted studies that suggest how the act of being persuaded is dependent on human aspects such as understanding, learning, context and work relations. Consequently, we argue that exploring how to organize usability work to include human perspectives and support cross-professional learning is a huge—but crucial—future challenge for work on downstream utility.

Dansk sammenfatning

Nærværende PhD-afhandling består af otte artikler, der undersøger usabilityarbejde i forbindelse med systemdesign. Disse perspektiver inkluderer evalueringspraksis, skriftlig feedback, tværfagligt samarbejde, og udfordringer for usability studier i computerspilsbranchen.

Artiklerne behandler tre primære emner: Usabilityarbejde forstået som en menneskelig aktivitet, usabilitypraksis og -metoder, og overbevisningskraft.

Usabilityarbejde forstået som en menneskelig aktivitet

Ofte er usabilityarbejde forstået som metoder og arbejdsprocedurer. I mine arbejder har jeg—inspireret af Naurs arbejde (Naur, 1992)—gentagende argumenteret for, at usabilityarbejde også handler om menneskelige aspekter som samarbejdsevne, forhandlingsvillighed, personlighed og humor (Nørgaard, 2007), og at en modning af vores forståelse af usabilityarbejde som en menneskeorienteret aktivitet vil bringe os tættere på en forståelse af, hvorfor usabilityarbejde ikke har haft den effekt på design, som praktikere og forskere kunne ønske sig (Nørgaard, 2007; Nørgaard & Hornbæk, 2008b).

Det menneskelige aspekt af usabilityarbejde vedrører både professionelle praktikere og systembrugere. Vores arbejde med evaluering af computerspil viser, at spilbranchen står med et usabilitybegreb, der ikke modsvarer de mest vigtige aspekter af computerspilsbrug (Nørgaard & Rau, 2007; Johansen, Nørgaard, & Sørensen, 2008). Brugsaspekter som udfordringsniveau og underholdningsværdi er på ingen måde dækket af det traditionelle usabilitybegreb, defineret af ISO 9241-11 standarden. Det betyder blandt andet, at de usabilityevalueringsmetoder, der er beskrevet i den traditionelle HCI-litteratur, ikke tilfredsstillende kan bruges til at evaluere computerspil. Af samme grund er der stigende opmærksomhed på at udvikle evalueringsmetoder, med særligt fokus på spil, se eksempelvis Fabricatore et al. (2002) og Medlock et al. (2002).

At usabilityarbejde er en menneskelig aktivitet har også betydning for, hvordan vi forstår de mennesker, der professionelt er involveret i systemdesign. Problemer med modstridende mål, prioriteter og personligheder er en stor udfordring for tværfagligt samarbejde (Nørgaard & Hornbæk, 2008b). Særlige karakteristika ved måden en organisation er opbygget på, kan ligeledes betyde jalousi og samarbejdsvanskeligheder (Nørgaard & Sørensen, 2008) mellem afdelinger, der for

eksempel begge mener at arbejde med user research.

I forbindelse med mit arbejde om usability som menneskelig aktivitet har to vinkler fået særlig opmærksomhed: læring og samarbejde. Fokus på, hvordan evalueringresultater bliver kommunikeret til designprocessen, synes at vise, at feedback, der tager udgangspunkt i at usability arbejde er en læringsproces, og lader udviklerne opleve eller erfare usabilityproblemer, tilsyneladende understøtter læring om usabilityproblemer bedre, end formater, der blot beskriver et problem med tekst (Nørgaard & Høegh, 2008; Nørgaard & Hornbæk, 2008a).

En lang række eksperter er involveret i systemdesign, og er dermed interessenter til usabilityarbejde. Af samme grund er samarbejde et vigtigt fokus, når man taler om det menneskelige perspektiv. Vores undersøgelser i industrien viser, at flere usabilitypraktikere har stort fokus på samarbejde, men at de gennem forskellige tiltag der skal støtte samarbejde—som for eksempel prototyping på tværfaglige workshops og aktiv inddragelse af udviklere i evalueringarbejdet—også ændrer deres egen jobfunktion (Nørgaard & Hornbæk, 2008b).

Samarbejde er, fra et organisationssynspunkt, interessant fordi forskellige arbejdsgrupper kan have overlappende arbejdsområder, hvilket—for eksempel når både en marketingsafdeling og en usabilityenhed laver brugerstudier—kan skabe grobund for rivaliseren og kamp om budgetmidler i stedet for at føre til et frugtbart samarbejde (Nørgaard & Sørensen, 2008).

Usabilitypraksis og metoder

Hvordan usabilityarbejde bliver udført i praksis, hvilke udfordringer praktikere slås med, og hvordan de forsøger at adressere disse, er andre temaer, som vores arbejder undersøger (Nørgaard & Hornbæk, 2008a; Nørgaard & Hornbæk, 2008b).

Litteraturen, der beskriver evalueringmetoder, synes ikke at hjælpe med at adressere alle de udfordringer, som praktikere møder i deres daglige arbejde. Vores studier viser, at der blandt andet er brug for fokus på, hvordan evalueringresultater kan analyseres systematisk. Der er ligeledes brug for et metodefokus på nogle af de opgaver—som for eksempel, hvorvidt brugeren oplever en applikation som æstetik eller sikker—som kunder stiller usabilitypraktikere, og som er svære at undersøge med for eksempel tænke højt protokollen (Nørgaard & Hornbæk, 2006).

Sådanne metodeproblemer er også til stede for user researchere, der arbejder med usability og user experience i computerspilsbranchen. Evalueringmetoder, som for eksempel tænke højt protokollen, kommer til kort, når opmærksomheden falder på brugsaspekter som underholdningsværdi og udfordringsniveau (Nørgaard & Rau, 2007; Johansen, Nørgaard, & Sørensen, 2008). Hvis vi skal lykkes med at modne usabilitybegrebet, bør udvikling af evalueringmetoder i fremtiden kunne rumme bredere brugsaspekter, som dem vi for eksempel kender fra spilverdenen.

Vores studier i industrien har peget på en række udfordringer der—selvom ikke alle er nyopdagelser—udgør en stor udfordring for samarbejdet omkring usability i system design. Dårlig timing af resultater, irrelevante resultater, mangel på respekt for kollegers faglighed, og vanskeligheder med at kommunikere (Nørgaard & Hornbæk, 2008b) er alvorlige problemer, der måske kan forstås bedre, hvis vi beskuer usabilityarbejde som et tværfagligt projekt der—udover at være bestemt af metoder og arbejdsprocesser—også er bestemt af menneskelige aspekter som hvordan mennesker lærer, tænker og samarbejder (Nørgaard, 2007).

Overbevisningskraft og det at blive overbevist

Flere af vores studier berører usabilityevalueringers overbevisningskraft, eller det fænomen/den proces, hvor for eksempel udviklere bliver overbevist om rigtigheden og relevansen af et usabilityproblem (Nørgaard & Høegh, 2008; Nørgaard & Hornbæk, 2008a). Downstream utility—altså det, at usabilityarbejde bliver brugt og implementeret i det efterfølgende designarbejde—er uden tvivl afhængig af metoder og arbejdsprocedurer, og måden hvorpå evalueringresultater bliver beskrevet og kommunikeret på er ligeledes væsentlig. De undersøgte måder at give feedback på, der rummer klare pædagogiske elementer såsom muligheden for at opleve brugsproblemer på egen krop, eller diskutere et brugsproblem på en tværfaglig workshop, syntes at være bedre til at skabe grundlaget for, at et usabilityproblem blev anerkendt, end formater, der primært hviler på en skriftlig beskrivelse af problemet (Nørgaard & Høegh, 2008). Samtidig er det vigtigt at understrege, at begrebet overbevisningskraft ikke må misforstås som en særlig kraft, der er indeholdt i for eksempel format X og ikke i format Z. Vores studier af måder at levere feedback på peger på, at mennesker ikke bliver overbevist af nogen særlig kraft, men at de konstruerer overbevisningen

på baggrund af et givent foreliggende materiale.
Dermed er det at blive overbevist om for eksempel

et usabilityproblem afhængig af læring, kontekst
og relationer, og dermed udpræget individuel.

Preface

This thesis is submitted in order to obtain the degree of Doctor of Philosophy at the University of Copenhagen, Faculty of Science, Department of Computer Science (DIKU). It is founded as part of the USE project, grant number 2106-04-0022 from the Danish Strategic Research Council, and work has been carried out between May 1st 2005 and May 1st 2008.

I wish to thank the many friends and colleagues who have helped me and inspired my work. First, my supervisor Kasper Hornbæk who has been a valuable critic and eager discussion partner, and whose knowledge of HCI seems ever-reliable, but also thanks to Erik Frøkjær and the rest of the HCI group at DIKU for their willingness to support, debate and share inspiration. I am also grateful to Alex Taylor and other colleagues at Microsoft Research in Cambridge, UK, whose eagerness to explore so diverse themes as human being, robots,

technology-enhanced human-pet interaction, and archaeology's relation to HCI have pointed to so many fascinating directions one could take HCI work in the future.

Janus Rau from IO Interactive has been a treasured colleague and guide in the world of computer games. A multitude of practitioners from the Danish industry has participated in my studies. You are too numerous to mention, but I am grateful for the time and effort you all have put into teaching me about your work.

Also thanks to fellow PhD fellows and friends Nicolai, Kristin, Julie, and Christina for valuable discussions of the hows and whys of research, colleagues and friends at Collaboration Lab for exercising my brain in new ways, and to my family, friends, and partner for simply putting up with me.

Introduction

This compilation of papers is a result of three years' work at Copenhagen University, Department of Computer Science where I have carried out my PhD studies in the HCI-group.

The papers discuss the themes usability evaluation and downstream utility from different angles, namely evaluation practice, persuasiveness of written feedback formats, cross-professional cooperation, and challenges for user research in the computer games industry. Figure 1 lists each paper, the research questions they explore, and their main results.

In order to sum up the scientific contributions of my work, I have examined abstracts and conclusions, identified major issues, and used affinity diagramming to condense three core themes in my work. Figure 2 relates themes and papers.

In the following—and before presenting the actual papers—I shall briefly reflect upon each theme.

Understanding usability work as a human activity

Usability work is often discussed in terms of

methods and work procedures. In our papers—in some more directly than others—we argue that usability work is much more, and that maturing our understanding of usability work to include a human perspective, is a step towards understanding why usability work have had less effect than intended by most of us who work with system design.

Let me shortly elaborate on what I mean by usability work being a human activity. The theme is inspired by the works of Naur (see for example Naur, 1992) who argued that computing is a human activity where human aspects such as understanding, learning and thinking play a crucial role for the success of for example programming. The same can be said about usability work. In our work that concern games development the human perspective is primarily turned towards the user. The view on users—and in particular users' reasons for using a system—in the games industry includes human aspects that differ from the traditional usability concept (Nørgaard & Rau, 2007; Hassenzahl & Tractinsky, 2006).

Traditional usability includes three pillars: effectiveness, efficiency, and satisfaction, which are all task-oriented, and where only satisfaction touches vaguely upon how humans experience a system.

ID	Title	Research questions	Main results
1	Nørgaard, M. & Hornbæk, K. (2006): What Do Usability Evaluators Do in Practice? An Explorative Study of Think-Aloud Testing, International Conference on Designing Interactive Systems (DIS 2006), June 26–28, University Park, Pennsylvania, USA.	How is think aloud testing currently practised in the industry?	<ul style="list-style-type: none"> • Descriptions of the Think Aloud protocol do not completely map the practical challenges for usability evaluations. • Results are rarely immediately analyzed, if at all. • Usability researchers encounter problems with investigating certain aspects of the system such as utility and general impressions
2	Nørgaard, M. (2007): Usability Work: A Human Activity. COST294-MAUSE workshop on downstream utility: the Good, the Bad, and the Utterly Useless Usability Evaluation Feedback, November 6th, Toulouse, France.	What might researchers gain from understanding usability work as something other than evaluation methods and work procedures?	<ul style="list-style-type: none"> • Thinking about usability work as a human activity—rather than something that is defined by methods and work procedures— may better reflect the nature and challenges of usability. • Understanding usability work as dependent on for example human’s ability to learn and collaborate may help researchers and practitioners understand and address the work challenges encountered by user researchers and other stakeholders.
3	Nørgaard, M. & Rau, J. (2007): User Testing in the Combat Zone. International Conference on Advances in Computer Entertainment Technology, June 13th-15th, Salzburg, Austria.	What are the specific challenges for usability and user research in the games industry?	<ul style="list-style-type: none"> • User researchers in the games industry encounter the same challenges as researchers in other industries, such as for example arguing for return of investment. In addition, they struggle with a usability concept and evaluation methods that do not fully cover the aim of their research.
4	Nørgaard, M. & Hornbæk, K. (2008): Working Together to Improve Usability: Challenges and Best Practices. Technical report from University of Copenhagen, Department of Computer Science, www.diku.dk/publikationer/tekniske_rapporter/rapporter/08-01.pdf	Which challenges do usability researchers, developers, and project managers encounter when they collaborate on usability work?	<ul style="list-style-type: none"> • Four main challenges to successful interaction between participant groups are identified: poor timing of usability results, results lacking relevance, little respect for other work disciplines, and difficulties sharing information. • Understanding usability as a cross-professional learning process helps explain the reasons and successful solutions for those challenges.

continues on next page

Figure 1 (this page and next): Overview of papers, the research questions they explore and their main results.

ID	Title	Research questions	Main results
5	Nørgaard, M. & Hornbæk, K. (2008): Exploring the Value of Usability Feedback Formats. <i>The International Journal of Human Computer Interaction</i> (in press).	How do different usability feedback formats perform in a use situation? Does the use and value of a feedback format change over time?	<ul style="list-style-type: none"> • Content-rich formats such as redesign proposals, screen dumps and multimedia presentations are initially favoured by developers over problem reports and scenarios. • After use developers rate all formats equally useful. • Feedback seemingly serves multiple purposes that change over time. First, it needs to convince developers about the relevance of a problem. Then, it must be easy to use in the daily work, and finally it must serve as a reminder of the problem.
6	Nørgaard, M. & Høegh, R. T. (2008): Evaluating Usability – Using models of Argumentation to Improve Persuasiveness of Usability Feedback. <i>The International Conference on Designing Interactive Systems (DIS2008)</i> , 25th–27th February, Cape Town, South Africa.	Can rhetoric models help explain the success and failure of feedback formats?	<ul style="list-style-type: none"> • Feedback that reflects the rhetoric models of Toulmin and Aristotle seems more persuasive than those that do not. • Aspects that relate to learning and pedagogy may better than theories of argumentation explain why some formats are considered more persuasive than others.
7	Johansen, S.A.; Nørgaard, M. & Rau, J. (2008): Can Eyetracking Boost Usability Evaluation of Computer Games? <i>CHI2008 workshop on Evaluating User Experiences in Games</i> , 4th April 2008, Firenze, Italy.	Can eye tracking boost usability evaluation of computer games?	<ul style="list-style-type: none"> • Eye tracking may support evaluation of attention-related aspects of gaming that are otherwise difficult to explore. • Heat maps and other tangible outputs from eye tracking may help user researchers argue for evaluation results and design changes
8	Nørgaard, M. & Sørensen, J.R. (2008): Organizational Challenges to User Research in the Video Game Industry: Overview and Advice, in Isbister, K. & Shaffer, N. (eds.) <i>Game Usability: Advice from the Experts for Advancing the Player Experience</i> , Morgan Kaufman.	What kind of usability challenge, unique for the games industry, arises from the organizational separation between developer and publisher?	<ul style="list-style-type: none"> • In the games industry, user research is often conducted simultaneously by different work groups, such as 3rd party developers and marketing departments. • The organizational divide between two groups of user researchers may pose unintentional competition between colleagues, jealousy and confusion about what user research is.

The 1998 edition of the ISO 9241-11 standard for usability describes satisfaction as 'Freedom from discomfort, and positive attitudes towards the use of the product' (International Organization for Standardization, 1998). In the games industry satisfaction is understood as comprising much more. Here, satisfaction is considered as being (also) influenced by emotions such as fun, challenge, curiosity and aesthetics (Nørgaard & Rau, 2007; Hassenzahl & Tractinsky, 2006; Monk, Hassenzahl, Blythe, & Reed, 2002). This way the computer games industry have put emphasis on users and usage as involving more than setting and reaching a series of work tasks without too much trouble.

Apart from how we understand users, the human perspective on usability work also impacts how we think about the professionals who are involved in systems design. Collaboration and communication among stakeholders to the development process have been subject of many studies (Bennet & Karat, 1994; Bødker & Buur, 2002; Bødker & Krogh, 2001; Hornbæk & Frøkjær, 2005; Madsen & Petersen, 1999; Uldall-Espersen & Frøkjær, 2007) some of which specifically discuss issues that refer to human relationships (Redish, Bias, Bailey, Molich, Dumas, & Spool, 2002; Bennet & Karat, 1994; Furniss, Blandford, & Curzon, 2007; Wixon & Wilson, 1997). In this respect cross-professional collaboration face challenges that arise from stakeholders having conflicting priorities, procedures and personalities (Nørgaard & Hornbæk, 2008b).

Certain organizational setups may also spur rivalry and jealousy between colleagues, as when two groups of professionals organized in, say, a marketing department and a user experience unit, both claim to do user research (Nørgaard & Sørensen, 2008) and perhaps fear losing influence or budget to the other party.

As a consequence, and if we are ever to address problems for usability work, such as rivalry and lack of respect, researchers need to understand that usability work is much more than methods and work procedures. That usability work is first and foremost a human activity.

While the term human perspective may comprise aspects such as creativity, personality, humour, social skills, ability to negotiate (Furniss, Blandford, & Curzon, 2007; Nørgaard, 2007; Nørgaard & Hornbæk, 2008b), and so forth, two aspects appear repeatedly in our work, namely collaboration and learning.

Learning about usability issues or problems is without discussion a crucial goal for usability

Theme	Papers
Usability work as a human activity	1,2,3,4,5,6,7,8
Practice and methods	1, 3, 4, 7, 8
Persuasiveness	5, 6, 7

Figure 2: The three core themes are related to the papers. For titles and other details, see Table 1.

work. In our papers this theme is closely linked to the feeding back of evaluation results to the on-going design process (Nørgaard & Hornbæk, 2008a; Nørgaard & Høegh, 2008). When trying to understand if certain feedback elements facilitate learning better than others, we found pedagogical aspects important for how well feedback was rated by developers (Nørgaard & Høegh, 2008). Feedback that lets developers construct their own understanding of a problem by, for example, letting them experience problems on their own or watch users struggle with a task, seemingly facilitate learning about usability issues better than mere descriptions of problems (Nørgaard & Høegh, 2008).

Usability work being a collaborative process, the issue of learning applies also to stakeholders. However, some stakeholders have trouble understanding or respecting other job roles than their own (Nørgaard & Hornbæk, 2008b). Developers, for example, have reported that user researchers do not understand crucial technical parts of the system in question and often provide useless evaluation feedback. At other times, the useless feedback is caused by vaguely described usability problems (Dumas, Molich, & Jeffries, 2004). With usability research being tuned increasingly towards downstream utility (Cockton, 2006), finding out how to organize usability work so as to support learning about other professions' goals and values is a huge challenge for the future. Securing balance between professional expertise and a broad understanding of the entire design process will not only be a challenge for practitioners but also for the people who educate practitioners to come.

With many experts involved in systems design, collaboration is another indisputably important angle on human perspective (Bennet & Karat, 1994). In terms of usability evaluation, collaboration may mean involving stakeholders in the preparation, user test, and analysis of results

(Coble, Karat, & Kahn, 1997; Kennedy, 1989; Dumas, 1989; Bennet & Karat, 1994; Bødker & Buur, 2002; Madsen & Petersen, 1999; Uldall-Espersen & Frøkjær, 2007; Schell, 1986; Nørgaard & Hornbæk, 2008b). In a case study a user researcher reports how such an approach has improved downstream utility, but also changed the role of the user researcher to include project management (Nørgaard & Hornbæk, 2008b). Two colleagues in user research report how they engage crucial stakeholders in cross professional design workshops where low fidelity prototypes are discussed and changed real time on site (Nørgaard & Hornbæk, 2008b). From an organizational view, collaboration may mean that a usability unit and a marketing department share information about what kinds of user research they plan, what their aims are, and agree on how they may assist each other. Thus, focus on collaboration may prevent groups or individuals rivalling or working in different directions.

Practice and methods

Part of our work deals with usability work in practice, the challenges researchers face, and how they tackle them. Literature on usability evaluation, such as Molich (2003) and Dumas & Redish (1993) seemingly does not reflect the challenges that practitioners meet in the industry. One of our studies show that evaluation results are rarely analysed after a test session, that issues such as utility are hardly investigated, and that user researchers are encouraged by customers to investigate overall impressions, feelings of trust and other issues, which are difficult to probe for when using the think aloud protocol (Nørgaard & Hornbæk, 2006).

While user researchers working with task oriented systems seem to struggle with making methods meet practical realities and demands, user researchers in the games industry also face other challenges. In some of our papers we discuss user research in the games industry, and argue that the concept of usability in games is not satisfactorily covered by for example the ISO 9241-11 definition (Nørgaard & Rau, 2007; Johansen, Nørgaard, & Sørensen, 2008). Games usability is a concept that must be developed beyond traditional usability, since concepts such as challenge and fun are crucial to games evaluation but not addressed by the traditional definition of usability (Hassenzahl & Tractinsky, 2006). Stuck with an ill-fitting usability definition, games researchers also suffer from not having a broad palette of evaluation methods to help them investigate for example a game's level of challenge. Though a great deal of

work has gone into developing evaluation methods for games (Malone, 1982; Medlock, Wixon, Terrano, Romero, & Fulton, 2002; Fabricatore & Rosas, 2002; Desurvire & Toth, 2004), there is still a long way to go before we understand how users experience games. Future work that aims to mature the concept of usability to fit games must therefore be followed by attempts to develop evaluation methods that focus on games-related aspects of usability such as fun and aesthetics. One can only speculate whether a fully developed games usability concept may in time reflect on traditional design and usability work, for inspiration see (Chao, 2001), turning usability researchers' attention to design aspects such as fun, challenge and aesthetics.

Through company visits and interviews with people working in systems design we have come to learn a great deal about the challenges for such work. Poor timing of evaluation results, results lacking relevance, little respect for other disciplines, and difficulties sharing important information are central challenges for the successful collaboration between user researchers, developers, and project managers (Nørgaard & Hornbæk, *Working Together to Improve Usability: Challenges and Best Practices*, 2008). Though these issues to some extent have been discussed previously (Rosenbaum, Rohn, & Humburg, 2000; Gulliksen, Boivie, Persson, Hektor, & Herluf, 2004), and thus should be well-known to researchers and practitioners alike, they remain a crucial challenge for how we organize and practice usability work. We have suggested understanding usability work as a cross-professional learning process (Nørgaard & Hornbæk, 2008b) to include these and other human perspectives (Nørgaard, 2007; Gulliksen, Boivie, & Göransson, 2006; Furniss, Blandford, & Curzon, 2007; Iivari, 2006) in usability practice and perhaps this way address these challenges that all relate to how humans think, learn, and collaborate.

Persuasiveness and the process of being persuaded

The final theme, that I want to draw attention to here, concerns persuasiveness, which—retrospectively—needs to be elaborated. The outset for working with persuasiveness was that the field has focussed a lot on developing and describing evaluation methods (Jeffries, Miller, Wharton, & Uyeda, 1991; John & Mashyna, 1997; Karat, Campbell, & Fiegel, 1992; Nielsen, 1992; Hertzum, 1999; Sawyer, Flanders, & Wixon, 1996), and less on how evaluation results may be communicated

successfully to stakeholders, though see (Dumas, Molich, & Jeffries, 2004; Hornbæk & Frøkjær, 2005). Downstream utility no doubt is related to evaluation methods and work procedures, and the way results are fed back to, say, developers, is absolutely crucial (Redish, Bias, Bailey, Molich, Dumas, & Spool, 2002). Since persuasiveness is not yet an established concept, and since I do not wish to detach it from my focus on the human perspective on usability work, I find that it useful to clarify the term.

Let me initially suggest persuasiveness as something that goes on in the human mind. To better reflect this, we should perhaps rather refer to persuasiveness as the process of being persuaded or simply being persuaded. This difference between noun and verb is important to make clear since our work on feedback formats (Nørgaard & Hornbæk, 2008a; Nørgaard & Høegh, 2008) may be misunderstood as understanding persuasiveness as an attribute of certain feedback formats and not of others. That format X is persuasive and format Z is not, for example. This is not our understanding. The process of being persuaded is much more dependent on human aspects such as understanding, learning, context and relations, and is thus highly individual.

To give an example, our attempts to understand being persuaded as dependent on how well feedback mapped well-known rhetorical models (Nørgaard & Høegh, 2008), gave no clear-cut answer. Rather, that particular exploration of feedback formats suggested that the process of being persuaded is perhaps best supported by feedback that rely on learning and pedagogy, such as self-experience or cross professional feedback workshops (Høegh, 2007; Nørgaard & Høegh, 2008).

To sum up, the three core themes of our work cover aspects of usability work related to practice and methods, to the maturing of the usability concept to better fit human aspects of user and usage, and to a human perspective such as for example thinking about stakeholders as learners and collaborators. Left is to ponder about which questions come next, and how one may go about exploring them.

Future work

To develop how we think about downstream utility, researchers need to address the problems of disrespect, irrelevance, poor timing, and poor communication which are still major challenges for usability work. In our work, we have seen promising examples with for example using inter-

disciplinary design workshops to involve stakeholders in user-centred design and evaluation. However, practitioners may end up struggling to find the balance between being experts on the one hand and multidisciplinary collaborators on the other. This practical challenge should not be let to each individual to deal with alone, but be thoroughly considered by the people who teach and influence future stakeholders to usability work.

Introducing the human perspective as a concept is only a very little step towards integrating understanding of human being, thinking and learning in usability and design work. As our studies of feedback approaches were not designed to explore human perspectives such as learning and understanding, one important challenge for future research is for example to explore how we may feed back evaluation results in ways that suits the individual, and in ways that are realistic within an organizational context. Moving away from thick usability reports, we might consider including for example stakeholders in the production of written feedback based on cross-disciplinary discussions of evaluation results. Further, with an increased focus on learning it may quickly become clear that one thing is learning about something, quite another is implementing it. Research that explores how we might facilitate this final step will be an important contribution to downstream utility.

Our look on computer games has proven the concept of usability too narrow to be useful, and games research has opened up for use perspectives closely related to the user's personal experience of for example fun and flow. Future work may look into how we can use this new perspective on users and system use in the context of task-oriented software. In terms of which roads to explore, focussing on fun and flow may be a valuable contribution to the design of traditional task-oriented systems

In the following I present the eight papers, first—and to provide an overview—in a collection of abstracts, then as complete papers.

References

- International Organization for Standardization. (1998). ISO 9241-11. International Standards for Business, Government, and Society.
- Bennet, J., & Karat, J. (1994). Facilitating effective HCI design meetings. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Celebrating Interdependence, 198-204.

- Bødker, S., & Buur, J. (2002). The design collaboratorium - a place for usability design. *ACM Transactions on Computer-Human Interaction (TOCHI)* .
- Bødker, S., & Krogh, P. M. (2001). The interactive design collaboratorium. *Proceedings of the Interact 2001* .
- Chao, D. (2001). Doom as an interface for process management. *Proceedings of the CHI 2001 Conference on Human Factors in Computing*, 152-157.
- Coble, J., Karat, J., & Kahn, M. (1997). Maintaining a focus on user requirements throughout the development of clinical workstation software. *Proceedings of the ACM Conference on Human Factors in Computing*, 170-177.
- Cockton, G. (2006). Focus, fit and fervour: Future factors beyond play with the interplay. *International Journal of Human-Computer Interaction*, 21, 2, 239-250.
- Desurvire, H. M., & Toth, J. (2004). Using heuristics to evaluate the playability of games. *CHI '04 extended abstracts on Human factors in computing systems*, 1509-1512 .
- Dumas, J. (1989). Stimulating change through usability testing. *SIGCHI Bulletin*, July 1989, 21, 1., 37-44.
- Dumas, J., & Redish, J. (1993). A practical guide to usability testing. Oregon, USA, Intellect Books.
- Dumas, J., Molich, R., & Jeffries, R. (2004). Business: Describing usability problems: Are we sending the right message? *Interactions*, 11, 4, 24-29.
- Fabricatore, C. M., & Rosas, R. (2002). Playability in Action Video Games: A Qualitative Design Model. *Human Computer Interaction*, 17, 4, 311-368.
- Furniss, D., Blandford, A., & Curzon, P. (2007). Usability Work in Professional Website Design: Insights From Practitioners' Perspectives. I E. Law, E. Hvannberg, & G. Cockton, *Maturing Usability: Quality in Software, Interaction and Value*, 144-167. Springer London.
- Gulliksen, J., Boivie, I., & Göransson, B. (2006). Usability Professionals - Current Practices and Future Development. *Interacting with Computers*, 18, 568-600.
- Gulliksen, J., Boivie, I., Persson, J., Hektor, A., & Herluf, L. (2004). Making a Difference - a Survey of the Usability Profession in Sweden. *Proceedings of Nordichi 2004*, 207-215.
- Hassenzahl, M., & Tractinsky, N. (2006). User experience - a research agenda. *Behaviour & Information Technology* , 25 (2), 91-97.
- Hertzum, M. (1999). User Testing in Industry: A Case Study of Laboratory, Workshop, and Field Tests. *Proc. ERCIM Workshop on User Interfaces for All* , 59-72.
- Hornbæk, K., & Frøkjær, E. (2005). Comparing usability problems and redesign proposals as input to practical systems development. *ACM Conference on Human Factors in Computing Systems*, 391-400.
- Høegh, R. (2007). Software Development and Feedback From Usability Evaluations. *Proceedings of ITAIS 2007, Venice, Italy* .
- Iivari, N. (2006). 'Representing the User' in Software Development - a Cultural Analysis of Usability Work in the Product Development Context. *Interacting with Computers*, 18, 635-664.
- Jeffries, R., Miller, J., Wharton, C., & Uyeda, K. (1991). User interface evaluation in the real world: A comparison of four techniques. *ACM Conference on Human Factors in Computing Systems*, 119-124.
- Johansen, S., Nørgaard, M., & Sørensen, J. (2008). Can eye tracking boost usability evaluation of computer games? *Workshop on Evaluating User Experiences in Games*, April 4th, CHI2008. Florence, Italy.
- John, B., & Mashyna, M. (1997). Evaluating a Multimedia Authoring Tool. *Journal of the American Society of Information Science*, 48, 9, 1004-1022.
- Karat, C., Campbell, R., & Fiegel, T. (1992). Comparison of Empirical Testing and Walkthrough Methods in Usability Interface Evaluation. *Proceedings of CHI'92*, 397-404.
- Kennedy, S. (1989). Using video in the BNR usability lab. *SIGCHI Bulletin*, 21, 2, 92-95.
- Madsen, K. H., & Petersen, M. G. (1999). Supporting collaboration in multi-media design. *Human-Computer Interaction - INTERACT'99*, 185-190.
- Malone, T. (1982). Heuristics for Designing Enjoyable User Interfaces: Lessons From Computer Games. *Proceedings of HumanFactors in Computer Systems*, Gaithersburg, Maryland, 63-68.
- Medlock, M., Wixon, D., Terrano, M., Romero, R., & Fulton, B. (2002). Using the RITE Method to Improve Products; a Definition and a Case Study. *Proceedings of Usability Professionals Association (UPA)*, Orlando, FL .
- Molich, R. (2003). Discount user testing. Hentet fra www.dialogdesign.dk
- Monk, A., Hassenzahl, M., Blythe, M., & Reed, D. (2002). Funology: designing enjoyment. *Workshop on CHI2002*, 924-925. Minneapolis, Minnesota, USA.
- Naur, P. (1992). Programming as Theory Building. I

- P. Naur, *Computing: A Human Activity*, 37-48. New York: Addison Wesley.
- Nielsen, J. (1992). Finding Usability Problems Through Heuristic Evaluation. *Proceedings of CHI'92*, 373-380.
- Nørgaard, M. (2007). *Usability Work - A Human Activity*. COST294-MAUSE Workshop - Downstream Utility: The Good, the Bad, and the Utterly Useless. Institute of Research in Informatics of Toulouse (IRIT), Toulouse, France.
- Nørgaard, M., & Hornbæk, K. (2006). What Do Usability Evaluators Do in Practice? An Explorative Study of Think-Aloud Testing. *Proceedings on the 6th ACM Conference on Designing Interactive Systems (DIS 2006)*. Penn State, Pennsylvania.
- Nørgaard, M., & Hornbæk, K. (2008a). Exploring the Value of Usability Feedback Formats. *International Journal of Human-Computer Interaction*, (in press).
- Nørgaard, M., & Hornbæk, K. (2008). Working Together to Improve Usability: Challenges and Best Practices. Technical report from Copenhagen University Dept. of Computer Science. <http://www.diku.dk/publikationer/tekniske.rapporter/rapporter/08-01.pdf>
- Nørgaard, M., & Høegh, R. T. (2008). Evaluating Usability - Using Rhetorical Models to Improve the Persuasiveness of Usability Feedback. *Proceedings of the 7th ACM Conference on Designing Interactive Systems (DIS2008)*, Cape Town, South Africa .
- Nørgaard, M., & Rau, J. (2007). User Testing in the Combat Zone. *Workshop on Methods for Evaluating Games - How to measure Usability and User Experience in Games*, The International Conference on Advances in Computer Entertainment Technology (ACE'07), Salzburg, Austria.
- Nørgaard, M., & Sørensen, J. (2008). Organizational Challenges to User Research in the Video Game Industry: Overview and Advice. In K. Isbister, N. Shaffer, K. Isbister, & N. Schaeffer (Red.), *Game Usability: Advice from the Experts for Advancing the Player Experience*, Morgan Kaufman.
- Redish, J., Bias, R., Bailey, R., Molich, R., Dumas, R., & Spool, J. (2002). Usability in practice: Formative usability evaluations - Evolution and revolution. *ACM Conference on Human Factors in Computing System*, Minneapolis, Minnesota, 885-890.
- Rosenbaum, S., Rohn, J. A., & Humburg, J. (2000). A toolkit for strategic usability: Results from workshops, panels and surveys. *Proceedings of the ACM CHI 2000 Conference on Human Factors in Computing Systems*, 1 , 337-344.
- Sawyer, P., Flanders, A., & Wixon, D. (1996). Making a Difference - The Impact of Inspections. *Proceedings of CHI'96* , 376-382.
- Schell, D. (1986). Usability testing of screen design: Beyond standards, principles, and guidelines. *Proceedings of the Human Factors Society 30th Meeting*, Santa Monica, CA , 1212-1215.
- Uldall-Espersen, T., & Frøkjær, E. (2007). Usability and software development: Roles of the stakeholders. *Proceedings of HCI2007*, July 22.-27., Beijing, China , 642-651.
- Wixon, D., & Wilson, C. (1997). The usability engineering framework for product design and evaluation. In M. Helander, T. Landauer, & P. P., *Handbook of Human Computer Interaction*, 653-688. North-Holland, Elsevier Science.

Papers

In the following the eight papers that this thesis builds upon, are presented. The papers are re-printed with the kind permission of the publishers.

[What Do Usability Evaluators Do in Practice? An Explorative Study of Think-Aloud Testing.](#)

Nørgaard, M. & Hornbæk, K. (2006) Proceedings on the 6th ACM Conference on Designing Interactive Systems (DIS'06), June 26–28, University Park, Pennsylvania, USA.

Think-aloud testing is a widely employed usability evaluation method, yet its use in practice is rarely studied. We report an explorative study of 14 think-aloud sessions, the audio recordings of which were examined in detail. The study shows that immediate analysis of observations made in the think-aloud sessions is done only sporadically, if at all. When testing, evaluators seem to seek confirmation of problems that they are already aware of. During testing, evaluators often ask users about their expectations and about hypothetical situations, rather than about experienced problems. In addition, evaluators learn much about the usability of the tested system but little about its utility. The study shows how practical realities rarely discussed in the literature on usability evaluation influence sessions. We discuss implications for usability researchers and professionals, including techniques for fast-paced analysis and tools for capturing observations during sessions.

[Usability Work: A Human Activity.](#)

Nørgaard, M. (2007) COST294-MAUSE workshop on downstream utility: the Good, the Bad, and the Utterly Useless Usability Evaluation Feedback, November 6th, Toulouse, France.

Much work on usability has a clear human perspective, such as making usability results more useful for developers. Yet, most work end up detaching usability work from human activities in its aim to isolate specific phenomena important to the quality and impact of evaluation results. This paper argues that researchers and practitioners could gain from understanding usability as a human activity involving, for example, learning about and understanding usability issues, and collaborating to improve usability.

[User Testing in the Combat Zone.](#)

Nørgaard, M. & Rau, J. (2007) Workshop on Methods for Evaluating Games - How to measure Usability and User Experience in Games, The International Conference on Advances in Computer Entertainment Technology (ACE'07), June 13-15, 2007, Salzburg, Austria.

This paper describes the how IO Interactive, a producer of computer games such as the Hitman series, has taken the first step towards working with us-

ability evaluations in a structured manner. The paper describes the usability team's first experiences with testing computer games and their work to integrate usability evaluation in the design of computer games. Finally, the paper identifies five categories of challenges that are vital for the usability team's success; justifying the costs of usability evaluation towards management; identifying structured work procedures that leaves room and opportunity for usability evaluation; identification and use of new methods to support the study of game-specific issues such as re-playability and game play; the ability to make alliances with important colleagues and managers; and identifying the people responsible for fixing usability issues.

Working Together to Improve Usability: Challenges and Best Practices

Nørgaard, M. & Hornbæk, K. (2008) Technical report from Copenhagen University Dept. of Computer Science, www.diku.dk/publikationer/tekniske.rapporter/rapporter/08-01.pdf

In theory, usability work is an important and well-integrated activity in developing software. In practice, however, collaboration on improving usability is ridden with challenges relating to conflicting professional goals, tight project schedules, and unclear usability findings. We study those challenges through 16 interviews with software developers, usability experts, and project managers. Four key challenges to successful interaction between stakeholders are identified: poor timing when delivering usability results, results lacking relevance, little respect for other disciplines, and difficulties sharing important information. We discuss practices that address these challenges, and present four guidelines to support the collaboration and professional relationship among developers, usability experts, and project managers. Our observations are further discussed as encompassing multiple perspectives and as a collaborative cross-professional learning process.

Exploring the Value of Usability Feedback Formats

Nørgaard, M. & Hornbæk, K. (2008) *The International Journal of Human Computer Interaction* (in press)

The format used to present feedback from usability evaluations to developers affects whether problems are understood, accepted, and fixed. Yet, little research has investigated which formats are the most

effective. We describe an explorative study where three developers assess 40 usability findings presented using five feedback formats. Our usability findings comprise 35 problems and 5 positive comments. Data suggest that feedback serves multiple purposes. Initially, feedback must convince developers about the relevance of a problem and convey an understanding of this. Feedback must next be easy to use and finally serve as a reminder of the problem. Prior to working with the feedback, developers rated redesign proposals, multimedia reports, and annotated screen dumps as more valuable than lists of problems, all of which were rated as more valuable than scenarios. After having spent some time working with the feedback to address the usability problems, there were no significant differences among the developers' ratings of the value of the different formats. This suggests that all of the formats may serve equally well as reminders in later stages of working with usability problems, but that redesign proposals, multimedia reports, and annotated screen dumps best address the initial feedback goals convincing developers that a usability problem exists and of conveying an understanding of the problem.

Evaluating Usability – Using Models of Argumentation to Improve Persuasiveness of Usability Feedback

Nørgaard, M. & Høegh, R. T. (2008) *Proceedings on the 7th ACM Conference on Designing Interactive Systems (DIS'08)*, February 25th-27th, Cape Town, South Africa.

Usability evaluation is widely accepted as a valuable activity in software development. However, how results effectively are fed back to developers is still a relatively unexplored area. We argue that usability feedback can be understood as an argument for a series of usability problems, and that basic concepts from argumentation theory can help us understand how to create persuasive feedback. We revisit two field studies on usability feedback to study if concepts from Toulmin's model for argumentation and Aristotle's modes of persuasion can explain why some feedback formats outperform others. We recommend that evaluators specifically back up the warrants behind their usability claims, that their arguments use several modes of persuasion, and that they present feedback in browsable amounts not to overwhelm developers with information. For complex and controversial problems, we advise evaluators to involve developers in a learning process and provide the opportunity to experience and discuss the findings.

Can Eyetracking Boost Usability Evaluation of Computer Games?

Johansen, S.A.; Nørgaard, M. & Sørensen, J.R. (2008) Workshop on Evaluating User Experiences in Games, April 4th 2008, CHI2008, Florence, Italy.

Good computer games need to be challenging while at the same time being easy to use. Accordingly, besides struggling with well known challenges for usability work, such as persuasiveness, the computer game industry also faces system-specific challenges, such as identifying methods that can provide data on players' attention during a game. This position paper discusses how eye tracking may address three core challenges faced by computer game producer IO Interactive in their on-going work to ensure games that are fun, usable, and challenging. These challenges are: (1) Persuading game designers about the relevance of usability results, (2) involving game designers in usability work, and (3) identifying methods that provide new data about user behaviour and experience.

Organizational Challenges to User Research in the Video Game Industry: Overview and Advice.

Nørgaard, M. & Sørensen, J.R. in Isbister, K. & Shaffer, N. (eds.) (2008) *Game Usability: Advice from the Experts for Advancing the Player Experience*, Morgan Kaufman.

In this chapter, we take a look at organizational challenges for 3rd party developers who are interested in implementing and conducting HCI-related user research, such as usability testing, in a game development setting. We discuss the challenges related to justifying the return of investment of user research, formalizing work procedures involving user research, and the building of cross-professional relationships amongst key stakeholders to user research. Furthermore, we also discuss the challenges related to the fact that many games developers are owned or closely affiliated with a publisher. Through the lenses of a questionnaire survey including members from the game industry, we specifically look at the relationship between 3rd party developers and the publisher's marketing department, and investigate how and to which extent these two parties collaborate on user research issues. During the chapter we also present concrete advice on how to tackle the various challenges mentioned.

What Do Usability Evaluators Do in Practice? An Explorative Study of Think-Aloud Testing¹

Mie Nørgaard

Department of Computer Science
University of Copenhagen
mien@diku.dk

Kasper Hornbæk

Department of Computer Science
University of Copenhagen
kash@diku.dk

Abstract

Think-aloud testing is a widely employed usability evaluation method, yet its use in practice is rarely studied. We report an explorative study of 14 think-aloud sessions, the audio recordings of which were examined in detail. The study shows that immediate analysis of observations made in the think-aloud sessions is done only sporadically, if at all. When testing, evaluators seem to seek confirmation of problems that they are already aware of. During testing, evaluators often ask users about their expectations and about hypothetical situations, rather than about experienced problems. In addition, evaluators learn much about the usability of the tested system but little about its utility. The study shows how practical realities rarely discussed in the literature on usability evaluation influence sessions. We discuss implications for usability researchers and professionals, including techniques for fast-paced analysis and tools for capturing observations during sessions.

Introduction

Methods for usability evaluation are one of the successes of human-computer interaction: they are widely used and in many cases improve the usability of the software to which they are applied. According to recent surveys (Gulliksen, Boivie, Persson, & Hektor, 2004; Vredenburg, Mao,

¹Originally published in Proceedings on the 6th ACM conference on Designing Interactive Systems (DIS'06), June 26th–28th, 2006, University Park, Pennsylvania, USA.

Smith, & Carey, 2002), think-aloud testing (TA) is widely used and valued by usability evaluators. Numerous studies have been made of usability evaluation methods in general, and of TA testing in particular (Hornbæk & Frøkjær, 2005; Jeffries, Miller, Wharton, & Uyeda, 1991; John & Mashyna, 1997; Karat, Campbell, & Fiegel, 1992; Nielsen, 1992); for recent reviews see (Cockton, Lavery, & Woolrych, 2003; Dumans, 2003). In our view, however, these studies are biased in two respects. First, most studies do not take place in a practical software development context, but in a laboratory-style set-up with non-expert participants. While such studies give insight into benefits and drawbacks of particular evaluation methods, they miss how practical realities of software development shape the use of evaluation methods (Wixon, 2003). Second, studies of usability evaluation tend to focus on coarse measures of outcomes such as the number of problems identified; they rarely describe the process of evaluation in detail. One exception is diary studies of usability evaluation, such as (John & Packer, 1995), which have provided valuable input on how evaluation methods are used. In a 2004 keynote, John called for more studies of the process of using HCI methods (John B., 2004), seemingly dissatisfied with the current literature.

Addressing the two biases above, this paper reports an explorative study of how TA testing is practiced. We do so by observing the setting up, carrying out, and handling of results from TA sessions in professional consultancies or software development organizations. Inspired by grounded theory and verbal protocol analysis, we analyze and summarize data with two expected benefits. For usability researchers, we intend the paper to deliver insights into some issues of practical usability work. For usability professionals, we identify some of the problems and tradeoffs they face, hoping that this may assist the planning and conducting of future TA tests.

Related work

The question of how TA testing is done in practice is related to studies (a) describing experiences from real-life usability evaluation or (b) presenting detailed information on the process of usability evaluation. Below we review this research and discuss the extent to which it helps understand the practice of TA testing.

One group of studies describes real-life usability evaluation. Some of these studies systematically collect data through observation and interviews of usability specialists and other stakeholders in

software development projects, see for example (Boivie, Åborg, Persson, & Löfberg, 2003; Iivari, 2005; Wilson, Bekker, Johnson, & Johnson, 1997). These studies focus on factors that facilitate or impede usability evaluations and the impact of their results. They have identified several strategic concerns in real-life usability evaluation, such as the need for users to be involved throughout the design process to facilitate useful contributions (Wilson, Bekker, Johnson, & Johnson, 1997) or that the organization of usability work, to some extent, shape usability results (Iivari, 2005). They do not, however, in detail discuss how evaluations are undertaken.

Other studies have focused more on tactical issues of usability evaluation, see for example (Dumas, Molich, & Jeffries, 2004; Hertzum, 1999; Molich, Ede, Kaasgaard, & Karyukin, 2004; Sawyer, Flanders, & Wixon, 1996; Szczur, 1994). These issues include how to make the results of usability evaluations such as TA testing impact software development (Hertzum, 1999; Sawyer, Flanders, & Wixon, 1996) and how to deliver feedback that is useful to developers (Dumas, Molich, & Jeffries, 2004; Hornbæk & Frøkjær, 2005). As an example, Molich et al. (Molich, Ede, Kaasgaard, & Karyukin, 2004) discussed how the usability reports produced by nine teams of mostly professional evaluators differ in content. They found great variation in selection of tasks for usability tests and in reporting of results. Studies of tactical issues of usability evaluation rarely describe the process but focus mainly on the outcome of usability evaluation.

Equally interesting are studies where professionals report how practical circumstances have forced them to adapt and develop the evaluation procedures they use, see for example (Arnowitz, Gray, Dorsch, Heidelberg, & Arent, 2005; Spencer, 2000; Zirkler & Ballman, 1994). Spencer (2000), for example, described how the evaluation technique cognitive walkthrough was modified to better fit the realities of the software development organization in which he worked. Those realities include time pressure and a defensive attitude among participants in the walkthrough. Spencer reported that the modified technique worked better in his organization. Such studies provide interesting observations on factors influencing practical usability work, such as the influence of a particular kind of product on the decisions about which evaluation method to use (Zirkler & Ballman, 1994). Yet, they lack the methodological rigor of the studies mentioned above and may not provide general lessons for usability research.

Another group of studies has focused on the process of usability evaluation. Mostly, the academic

literature on usability evaluation has been concerned with the outcome of evaluation in the form of problem lists or suggestions for redesigns. A few studies, however, have reported diary studies of usability evaluation (Hornbæk & Frøkjær, 2004; Jacobsen & John, 2000; John & Packer, 1995). In those studies, evaluators typically keep a diary in which they make notes on their planning, conducting and reporting of an evaluation. John and Packer (1995) showed how participants in a diary study made severity judgments based on personal judgment rather than on the usability evaluation technique used. Hornbæk and Frøkjær (2004) argued that the evaluation process observed in their diary study was complex, with participants identifying usability problems not just while conducting the actual evaluation, but also during planning and reporting of the evaluation. The studies referenced above, however, look only at non-expert evaluators outside an industrial software development context. These studies, and studies where the evaluator fill out forms during evaluation (Cockton, Woolrych, Hall, & Hindmarch, 2003), present the most detailed data on evaluation currently available. We know of no studies that have systematically observed and analyzed usability evaluation, for example using video. Overall, it appears that studies looking at real-life usability evaluation place little focus on describing the process of usability evaluation; studies of the evaluation process look at somewhat artificial evaluation settings with diaries as the data-collection method with the finest granularity.

The paper by Boren and Ramey (Boren & Ramey, 2000) is a notable exception to these shortcomings. Boren and Ramey observed TA sessions in two companies, and related their observations to what some consider the theoretical basis of TA testing, the work of Ericsson and Simon (Ericsson & Simon, 1993). The analysis by Boren and Ramey showed discrepancies between the observed TA testing and the recommendations of Ericsson and Simon. While the work of Boren and Ramey has given unique insights to usability research, it is limited in that they reported mainly discrepancies to Ericsson and Simon's prescriptions (in particular about prompting the user), and not more general issues confronting a usability specialist conducting an evaluation.

Attempting to broaden the focus of Boren and Ramey's paper we next present an explorative study concerning how usability evaluations are conducted in practice.

Exploring the use of think-aloud protocol

The question guiding the study is: what do usability evaluators do in practice? To get a better understanding of this we observed 14 TA test sessions in seven companies. We chose to focus on TA testing because it is widely used and because observing analytic usability evaluation, such as heuristic evaluation, presents methodological difficulties (e.g., concerning introspection) that we wanted to avoid. Our data comprise mainly audio recordings of the setting up, running and analysis of the TA sessions. Our intention is not to reprehend the practice of usability testing. Rather, we aim to explore what usability evaluators do so as to (a) sensitize usability research to industrial practice and (b) help evaluators understand better the strengths and weaknesses of what they do.

Companies Participating in the Study

Seven companies agreed to participate in the study by letting us observe how they conduct TA tests. The companies were recruited among Danish enterprises that either offer usability evaluation as consultancy or integrate usability evaluation in their systems development. Table 1 provides a summary of the companies; their names replaced by the letters A through G.

Our sample comprises three companies that provide usability evaluations solely to customers outside of the company and work with information technology as part of their core business (companies B, D, F). Two of the companies in the sample (companies A, C) perform usability evaluation both in-house and to customers outside of the company. These two companies have information technology and systems development as their core business. Finally, two of the companies solely perform usability evaluation in-house (companies E, G); while both companies have a strong presence online, their core business is in the service sector. The companies vary in size from 2 to 8500. They had varying levels of experience with usability evaluation; some of the evaluators we observed had only worked with usability for one year, while one had been conducting usability evaluations for eight years. Four companies evaluated running prototypes (companies A, C, E, F), two companies evaluated deployed applications (companies B, D), and company G evaluated paper prototypes. All tests observed were formative tests in that they were usability evaluations with users seeking to investigate issues such as concept, tools and navigation.

Company	A	B	C	D	E	F	G
Employees (working with usability)	810(6)	2(2)	165(3)	7(7)	8500(7)	16(3)	3464(8)
Test sessions observed	2	1	1	1	4	3	2
Evaluators present during tests	2	2	2	2	2	2	1
Evaluators' experience in years	1-6	1-8	2.5-6	1-6	4-6	1.5-6	6.5
Customer of test results	Intern	Extern	Intern	Extern	Intern	Extern	Intern

Table 1. The companies participating in the study and the test sessions observed within each company.

Data Collection

Methodologically we were inspired by grounded theory which dictates that researchers should not initiate an investigation on the basis of a list of hypotheses (Pace, 2004). Our data collection was thus broad and open-ended. We tried to participate in as many of the activities surrounding the usability evaluations as possible, wanting to probe how the TA protocol is put into practice. Data was collected over a period of three months and the focus of attention developed during this time, as suggested by (Strauss & Corbin, 1998; Pace, 2004).

The core of our data is the observations, field notes, and audio recordings from 14 TA sessions, that is, the period of time from the arrival of the test participant until that participant leaves. These sessions were distributed among the companies as shown in Table 1; the number of sessions we could observe was largely dictated by practical circumstances. In all sessions, except those of company G, two evaluators from the company were present. On average, an evaluation consisted of a series of six sessions, of which we typically participated in two. The sessions we participated in were placed both at the beginning, middle and end of the series. In one session, the recording made from an observation room was of such poor quality that it allowed only sporadic transcription of the interaction between user and evaluator.

When possible, discussions, analysis, and informal conversations among usability evaluators

before and after the test sessions were also observed and recorded. Sometimes customers (i.e., the persons who commissioned the test) were also present and took part in these discussions (e.g., company B). In two cases we recorded when usability evaluators delivered test results to the customers (companies C and G). In two cases we collected reports, summaries or notes that documented the tests (companies F, G). In two cases (companies A and C) we additionally conducted semi-structured interviews with the persons responsible for the usability work in the company.

The data collection described above resulted in, among other material, 24 hours and 54 minutes of audio recordings. Below we focus on the test sessions and the discussions immediately following tests—we only mention material from feedback sessions, usability reports, and the semi-structured interviews, when it corroborates findings from the core data.

Data Analysis

Analysis was conducted in three phases. First we segmented the recordings applying descriptive keywords to each segment. Second we re-evaluated segments and keywords in order to adjust keywords or apply new ones. Third we analyzed and tried to form a coherent interpretation of segments that shared keywords. We explain this procedure more thoroughly below, and briefly relate it to grounded theory (Pace, 2004) and Chi's proposal for how to analyze verbal protocols (Chi, 1997).

Segmenting and open coding of the recordings

The audio recordings were initially divided into 641 segments. One segment could concern a usability evaluator analyzing the test results, or explaining how to ensure scientifically valid test results. A segment could last from a few seconds to several minutes. We chose to do only a partial transcription of the recordings, but listened repeatedly to the segments during our analysis.

In order to code the segments, keywords were attached to each segment allowing us to analyze and group segments. Thirty-five keywords were generated as the study proceeded. Some segments regarded more than one interesting topic and hence got more keywords attached to it. This process is similar to open coding in grounded theory (Pace, 2004) or to Chi's (Chi, 1997) phase of developing or choosing a coding scheme or formalism.

Re-evaluating and crosschecking the coding

In order to ensure that a segment contained evidence for a specific keyword, the coding was carried out in two iterations, one by each of the authors. Disagreements or questions about the attachment of a keyword to a segment were discussed before attaching an existing or creating a new keyword. This is similar to Chi's phase of operationalizing evidence in the protocols (Chi, 1997) and, in part, to axial coding in grounded theory (Pace, 2004).

Synthesizing and interpreting the data

Groups of segments, which shared the same keyword, were analyzed to identify the most interesting areas and thus reduce the size of data. For interesting areas, we looked for the observations that were most surprising to us, or seemed to contrast the literature on usability research and textbook recommendations on how to do a usability evaluation. Such areas were selected for further analysis and interpretation. This phase is similar to Chi's phases of seeking patterns in the mapped formalism (Chi, 1997) or selective coding in grounded theory (Pace, 2004).

Results

The following section describes our results organized in six areas. Table 2 summarizes these areas and the main findings within each of them.

The areas concern (1) analysis of the results from a session, (2) confirmation of known issues, (3) practical realities, (4) questions asked during a test, (5) laboratory-style scientific standards, and (6) uncovering usability problems or utility concerns. Below we present each area in turn. For

findings we give the number of sessions in which they were observed. We use sessions rather than segments as an indication of frequency, because the number of segments is strongly influenced by the nature of a session, especially how much the evaluator and the user talks, how much they jump between topics, etc.

Analysis of results from a test session

The first area concerns how usability evaluators analyze test sessions. By analysis we mean the task of understanding and agreeing upon important observations from a session. Analysis also includes attempts to understand the causes of those observations, interpret user behaviour and find design solutions to observed problems.

None of the sessions included attempts to carry out a structured analysis of the results immediately after the session, for example by systematically agreeing on and then analyzing, say, the ten most prominent observations of user difficulties. However, as we have not in this study covered every step from test design to final report, we are not able to say if analysis took place later.

One evaluator did carry out a semi-structured analysis in the last minutes of three sessions though, focusing on summarizing key findings while the user was present:

F1: "Let's sum up: The front page [should] maybe emphasize what they have in mind [...] and the logo [gesturing where a logo should be]....and eventually [we should] list these sections. And the picture behind [we should] make it a bit more interesting. The editorial ends down here [points]..."

Three other evaluators (in a total of five sessions) also tried to sum up a few problematic topics and return to those topics for further questioning before ending the session. However, we did not encounter any systematic attempt to cover the most important observations directly after a session.

After a session had finished, the most common activity was that usability evaluators, and in four sessions also customers, discussed the session. We observed how they presented overall impressions intertwined with a general discussion about the system, social talk, observations, ideas for redesign, and occasionally analysis of the problems. To illustrate, an 11 minutes long discussion of a session was shaped as follows:

Impressions of user attitude, discussing problems with prototype (3 min); Identification of one problem, analysis, summary of observations from session (2 min); Talk about old ideas, identify two problems, analysis (2 min); Discussion of recom-

Areas of attention	Main finding	N	Quotes and examples of observations
Analysis of results from a test session	Analysis is unstructured	9	Scattered fragments of analysis; no systematic approach used
	Analysis is incomplete	9	Does not identify causes or solutions; restricts discussion to user traits
	Analysis as a summary with the user	3	“Let’s sum up”, selecting a few problems for further questioning; listing key findings
Confirmation of known issues as a test’ focus	Looking for known issues	8	“Now, I am just looking for ammunition”; develops ideas of problems before testing; tasks and questions designed to point out known issues
	Practitioners have foreseen problems	5	“We have a gut feeling”, “I told you so ”
Practical realities influencing tests	Technical problems	8	System breaking down; long response times in test environment; installation or security messages interrupt workflow
	Unfinished prototypes	6	Parts of prototype missing or inaccessible; “a log-in name should not be WaddleFish”; texts and pictures are wrong or out of date
Questions asked during a test	Problems are explained, not experienced	13	“Do you think you would go back to the front page?”; “did you notice this column?”; “what do you expect to see?”
	Leading questions	13	Questions address certain parts of GUI or system; evaluator hints the solution; “Can you do this another way?”
	Unnecessary or obvious questions	10	“You did figure out to press the print button?”; asking user to locate information that clearly appear on the present screen; asking if user would like relevant information
Trying to meet laboratory-style scientific standards	Evaluators want similar conditions for users under test	5	“We have to make sure all users get the same questions”
	Rigid or artificial procedures	3	Laboratory-style procedures; Danish evaluators speaking English to a Danish user; measurng subjective satisfaction overly systematic
Uncovering usability problems or utility concerns	User points to utility or lack thereof	10	“I would not like this”; user chooses to solve task without help of system
	Evaluator probes utility concerns	7	Asking about normal workflow; asking whether a task is realistic; “What would you typically do?”

Table 2. Overview of results. N refers to the number of sessions in which a finding was made (out of 14 sessions in total).

mentations and re-design (30 sec); Customer calls—and gets a short general summary (1.5 min); Summary of findings combined with general talk (2.5 min).

After this discussion, one evaluator went on to write a summary of findings to the customer. In other sessions the evaluators would just have a short conversation about general impressions before leaving the room, and thus ending the attempt to carry out an immediate analysis.

In nine sessions we saw examples of incomplete analysis. By incomplete analysis we mean remarks or observations that, if they were intended to assist in uncovering usability problems and solutions to such problems, needed to be elaborated and discussed. In seven sessions, for instance, evaluators would quickly characterize a user as being for example confused or insecure, but fail to follow up on this characterization or even identify what made the user become confused or insecure.

Confirmation of Known Issues as a Test Focus

The evaluators made comments before, during and after sessions, which let us to believe that they held more or less strong ideas about usability problems of the particular system being tested, even before commencing on the test. These ideas appear to shape the design of tasks and the questions raised during a test session. While such ideas are natural and may be important hypotheses, they sometime appear to focus the test on a particular topic or hypothesis. This delicate balance seem difficult to master.

After a session one evaluator stated, for example, that the test should provide proof for the conclusions in a usability report, which she had already begun writing:

C1: "I think we agree on many of the issues"

C2: "Yes – I have already written the chapter, I just need the ammunition".

A total of four evaluators stated that they had a more or less clear idea of the usability problems before commencing a test. In an interview another evaluator said that usability tests in some cases merely serve to confirm the evaluators' assumptions:

A: "When we design a test we practically always have a gut feeling where it will fail [...] in a way it is just an 'I told you so'-kind of thing, but it is nice to be able to document it".

The quotes suggest that usability evaluators see a need to support expert opinion with something more concrete when presenting customers with advice on usability. This may lead to tests that in part serve only to confirm.

In addition to these expressed opinions, it also appears that the actual activities of a test are sometimes chosen to confirm, or at least explore, areas known to be problematic. Questions and tasks within a test, for example, would be chosen to explore well-known issues. This led to test situations where evaluators literally waited for the user to point to the problem area. A1 explained to us how a certain task that required the entry of percentages most likely would cause problems. During the test, the user did actually spot the problem, and the response from the evaluator suggested almost a relief that the user did so:

U: [Typing]

A1: "So you just added minus 10 on both lines?"

U: "...And then I got 20%....WHAT?"

A1: "Yes" [laughs out confirmingly]

In another session, in response to a user severely criticizing a particular functionality, the usability evaluator broke out in laughter and said "this is really good", suggesting to us, that this issue was already anticipated as being problematic. In this way, 8 of the 14 sessions had examples of evaluators directly or indirectly expressing that they were confirmed in their preconceived opinions about usability problems.

It is hard to say whether a test focused at confirmation influences how evaluators interpret the observations they make during a test. An evaluator from company A noted after a session, "we really wanted to test this because we are confident it will fail, he [the participant] managed it, but I am sure others will not". The quote suggests that the expectation to find the problem in future tests could overshadow the possible interesting observation that at least one user successfully used a particular part of the interface. We return to discuss the balance between known issues and new findings in the discussion.

Practical Realities Influencing Tests

The study revealed numerous practical problems that usability evaluators experience when testing. In 12 sessions we observed examples of such problems or practical realities. These include system failures, users not showing up for a session, disturbing surroundings, and technical problems with recording devices. Despite such problems the evaluators managed to carry out all of the ses-

sions.

Data show that the practical realities surrounding a test are produced by many factors, some out of the evaluators' control. In eight sessions, for example, we observed severe technical problems interfering with the session. As an example one session had a technical problem approximately every five minutes, each resulting in a break in workflow.

In two sessions problems arose because the customer had failed to provide the required number of test participants, thus forcing the evaluators to quickly find a solution in order to carry through the test within the scheduled time:

F1: "The next user is one of my old friends [...]"

F2: "[...] they are not the first ones we choose, but if the customer fail to recruit [when they have agreed to do so] then we take whomever we can get."

Six sessions had problems with unfinished prototypes or last-minute changes to the prototype. One evaluator noted:

D1: "Some things will, if not done properly, affect the users' perception rather dramatically...A log-in name should not be "Waddle-Fish", it's such a developer-kind-of-thing to make up funny log-in names like that"

Unfinished prototypes or prototypes recently changed are two reasons that evaluators often were confused or in doubt about the functionality of the prototype. In seven sessions evaluators stated that they were not familiar with aspects of the prototype's functionality:

G1: "Now...let us see...[searches in prototype paper sheets]...these are brand new, so I have not looked at them before"

In sum, severe practical problems in some sessions lead to a continual interruption of the participants' attempts to complete their tasks. In this study, the practical realities influencing tests are much more frequent and severe than one would expect from textbooks or research papers on usability evaluation.

Questions Asked During a Test

The study showed variations in the types of questions asked by the evaluators. We analyzed these to understand which kinds of information usability evaluators are interested in, and to discuss later the validity of the information gained by different kinds of questions.

A large number of questions were reminders to keep talking like "Hmmm" and "Yes?". These kinds of questions were omnipresent and should be uncontroversial. Equally unsurprising is the many questions that simply try to elicit what the user is currently doing, or what problems the user is facing, for example "What is happening?", "What are you looking for?", or "What is the problem?". Many of these questions concerned the users' experienced problems in solving concrete tasks.

We encountered evaluators asking questions that differed dramatically from how Ericsson and Simon (Ericsson & Simon, 1993), and in part also Boren and Ramey (Boren & Ramey, 2000), suggest to interact with test participants. Some questions concerned, for example, nonexistent parts of the system, such as asking how the user would use a mouse to interact with a paper prototype or what the user would feel about having to create a user profile in order to be able to use the system.

Other questions appeared speculative or hypothetical. One evaluator asked, for example, "Do you think you would go back to the front page at some point?" and "Let us say that something here [in a list of articles] would interest you..." (both F1), asking the user to continue on this assumption.

Some questions urge users to look back in time and remember their thoughts, that is, retrospective questions. For example "Did you notice this column [when you were here before]?" (F1), or "Do you remember if you got what you expected from the web shop?" (E1).

Questions about the user's expectation of the system were also frequent, for example: "What would you expect to see?" or "How many would you expect to find?" (both from company G). Questions about the expectations of the system were often asked in the beginning of the session, for example:

D1: "Then you enter this page, and my first question is: Try looking at the page and try not to click on anything but just tell me what is happening on this page, what can you do, how do you like it and give me all of you general impressions. You may go into detail and if you point at something you are encouraged to do so with the mouse so that the secretary can see what is going on"

Another type of question apparently aims to elicit information about the users' feelings, typically by asking directly about what the user liked, trusted or were interested in. E1 asked, for example "So... you feel more secure now...or?", and F1 probed "Is there anything where you think: 'Wow! I would

like to click on that'...or?'"

In 13 sessions we observed one or more questions of the five kinds described above. In contrast to the experienced problems discussed earlier they did not concern problems experienced as part of solving a task, but rather imagined, indirectly experienced or expected problems. This intensive probing for such problems surprised us.

Thirteen sessions showed another kind of question, best characterized as leading questions. One evaluator, for instance, asked a question aiming at a certain issue of interest and the user would without much trouble solve the task or answer the question as anticipated:

[The user has pressed play to see an episode of a series of video clips in a media player:]

G1: "What would happen when this episode was over?"

User: "The series would end"

G1: "It was just a short version of the series or what?"

U: "[...]I have pressed to see the whole series...Ah! I have pressed to see the whole series [...] something [other episodes] could come afterwards [...]"

Trying to Meet Laboratory-Style Scientific Standards

The evaluators made several remarks suggesting that they find validity to be of great importance when testing. The concepts of validity upon which evaluators rely seem primarily to be those of scientific experiments, such as keeping the same procedure throughout a test, using representative subjects, and using elaborate questionnaires to get information on users' satisfaction. Note that we here mainly describe the evaluators' beliefs; in the discussion we will look closer at the relation of these views to those presented by the literature.

Evaluators from three companies (representing five sessions) emphasized that one should not change a test design between the sessions of a test. Changing a test design could include making changes to questions, tasks, prototype and choice of language. One evaluator, for example, stated the importance of maintaining the same tasks and phrasings of questions throughout a TA test even though it was evident after a few test sessions that the users misunderstood some of the tasks.

C1: "I think it is really annoying that we already now can see problems, which we cannot correct as we go along...but we have to

make sure that all users get the same questions"

In three sessions we observed how the fact that evaluators were trying to adhere to laboratory-style validity resulted in rigid and artificial procedures. For instance, we observed a session where Danish evaluators asked questions in English to a Danish user. The aim was to make test conditions similar among Scandinavian participants. In another session, evaluators tried to collect data about the system through a series of questions (e.g., "I will be more effective with the system") that users should rate on a one-to-seven scale. These questions are similar to instruments for measuring subjective satisfaction typically used in laboratory-style experiments. While such scales certainly have their uses, in this case they seemed to contradict what had happened during the session minutes before. This observation was supported by the evaluator:

A1: "When users rate statements [...] we take the results with a kilo of salt. This guy—it is a pretty good score right but [...] in the beginning he was right-clicking all over the place and he mentioned that he did not like the buttons disappearing..."

Thus, the questions were seemingly included to adhere to some perception of how scientific user testing should be conducted. In this case, the answers were apparently not used, but had they been, it might have led to a de-emphasis of the user difficulties just observed.

In sum, the attempt to adhere to scientific standards in some cases lead to rigid or artificial procedures that appeared unnecessary given the influence of practical realities and the rather informal analysis of test results mentioned earlier.

Uncovering Usability Problems or Utility Concerns

All sessions in the study would naturally include segments where usability problems were identified, including problems with scrolling, positioning of information, how links should be emphasized, how the user was prompted for information several times, etc. Other segments concern the utility of the system, for example which tasks the system should support or whether tasks from the test were unrealistic with regard to how the user usually uses the system or would want to use the system (Nielsen, 1993). We observed utility concerns being discussed in 10 sessions.

In seven sessions we observed how the evaluator asked more or less specific questions concerning the utility of the system. Consider the following example:

F1: "Lets look at the article again...What would you typically do?"

User: "I would pass it on...if it was fun and interesting..."

F1: "Like printing it?"

U: "No just by word of mouth..."

F1: "Word of mouth. Ok..."

U: "...unless it was really good - then I would forward it electronically..."

F1: "Would you ever print articles?"

U: "No...I actually save them [...]"

F1: "So...do you copy the text and paste it into a Word document?"

U: "Yes, I could do that"

Ten sessions had examples of users who were pointing to utility problems like the following from company C:

U: "[reads question loud:] 'Where would I look for an employee?'... I would use a phone-book [which is not a part of the system]"

Some users specifically pointed to areas of the system, which they found failed to support their workflow, for example from company E: "This is just to tell you that I would not do it like this".

In 13 sessions we observed how problems relating to usability seemed to be favoured over problems relating to the utility of the system. A remark from a user about not wanting to solve a task in the way suggested by the system did for example not result in an attempt to investigate that utility problem further; nor did it get reported to the customer during the feedback session we observed. This study suggests that utility problems are much less frequently examined than usability problems. Given the little attention problems regarding utility got in the sessions we observed, we do not expect them to be treated more thoroughly in discussions that we did not attend.

Discussion

To sum up, this study shows that careful and systematic analysis of usability problems rarely take place immediately after the sessions in which they occur. Evaluators do not always, either, ensure that they agree on even the most important observations from a test. In addition, many tests appear to search also—and sometimes mainly—for confirmation of issues known beforehand or observed in other tests. Most of the sessions we observed were affected by practical realities such

as incomplete prototypes and evaluators' limited experience with the system being tested. The questions raised by the evaluators during the test varied, but some questions appeared hypothetical and probed only users' expectations and not the problems they actually experienced. Some evaluators seemed to regard TA testing as a scientific laboratory-style method resulting in rigid and artificial procedures when conducting the test. Finally, seemingly important observations about the utility of the system being evaluated were made during sessions. These were infrequent, however, compared to results and discussions concerning usability issues.

Most surprising to us is the lack of systematic analysis while the results of a test are still fresh in mind. As we have not covered every step from test design to final report in this study, we are unable to rule out whether analysis was done at a later stage. Still, the fact that evaluators rarely check whether they agree on the most important observations from a session adds to the picture of analysis as being a weak part of the evaluation process. Work on the evaluator effect (Hertzum & Jacobsen, 2001) show that evaluators observing the same test find substantially different usability problems, making collecting and discussing different views of the main observations important. Summaries of the main observations by the evaluator while the test participant was present worked well—similarly to the idea of cooperative usability testing (Hornbæk & Frøkjær, 2005). However, using this or similar techniques to agree on observations from a test does not in itself reveal usability problems, the causes of those problems, or possible remedies for them.

Perhaps the lack of systematic analysis is understandable, given the scarce advice about analysis of usability tests we receive from textbooks and introductions about how to do a TA study. Molich (Molich, 2003), for example, used 2 pages of his 33-page instruction on how to do TA testing to discuss analysis. Dumas and Redish (Dumas & Redish, 1999) used around 31 pages of their 404-page textbook on analysis. It appears desirable that usability research develops and validates techniques supporting fast-paced analysis. Usability evaluators would be well advised to more systematically relate and discuss their observations when they are fresh in mind. Evaluators might take up using post-it notes for capturing observations during a session, and analyze these immediately after the session. They might find it rewarding to prioritize these post-its, possibly together with the user, to develop a common understanding, and discuss problems and feasible solutions.

The extent to which usability practitioners already before testing had a clear idea of the usability problems to be found was surprising. Interestingly, recommendations are made in the literature e.g., (Dumas & Redish, 1999, s. 160) about looking for known problems. Some views of the psychology of confirmation suggest that as a result of this, evaluators are very likely to confirm what they are looking for, perhaps failing to make other equally important observations. If the answer is not known with confidence prior to testing, we agree with the practice of exploring these explicit questions in the test. However, if usability issues are already known with such confidence that the practitioner is only “looking for ammunition”, why test at all? Finding the balance between on the one hand testing specific areas of concern and on the other hand exploring the system in a more open manner seems to be an important but difficult challenge to evaluators.

The practical realities surrounding the tests we observed are far from the expectations about the test situation presented in textbooks such as (Dumas & Redish, 1999). Techniques and tools that are usable under such less-than-ideal circumstances are needed, for example to enable the analysis of observations in the usually short time available between sessions. Evaluators should for their part consider preparing material to be used on the fly in case of system failure.

Given the work of Boren and Ramey (2000), we had expected open and varied questions. Quite surprisingly we saw hypothetical questions, abstract questions, leading questions, and plain impossible-to-answer questions: in short, questions that did not aim at understanding problems experienced by the user, but rather at encouraging users to predict possible problems. On the one hand this suggests that evaluators may be looking for information about feelings and perceptions, which cannot be gained from a traditional TA testing. On the other hand we feel obliged to point out that some of the questions we encountered could never produce useful answers.

Questions about “first impressions”, “what would you expect to be there [e.g., on the next page]”, or “what do you feel about this” may imply that evaluators need researchers to provide more valid and systematic ways of probing for, say, participants’ feelings of trust. Evaluators are advised to pay closer attention to the way they phrase their questions.

Questions probing for information about utility also seem to warrant further investigation. Molich (Molich, 2003) suggested asking test participants about their impressions of the tasks after a TA session. In two sessions we observed how

useful discussions about the users’ real-life tasks developed from such a question being asked during a test session. However when the same type of question appeared at the end of a session as advised by Molich, it became more general and received also a general answer. We suggest for researchers to provide further techniques for initiating discussions about utility during tests, which would help address the concern that usability testing might “tune a user interface at the tail end of design, to clean up any rough edges or unnecessary difficulty in understanding or interacting with the interface” (Beyer & Holtzblatt, 1998, s. 373), instead of concern the user’ tasks or needs. In order to understand and discuss how to improve the utility of a system evaluators may find it helpful to question the system’s utility and ask users how they usually go about solving a specific task.

The study suggests a belief amongst some evaluators that usability testing is science, and therefore must meet the same criteria as science. Iivari (2005) recently reported an explorative study in which similar attitudes were present among some usability professionals, “staid researchers” in Iivari’s terms. The insistence on, for example, not changing tasks or procedure during a test appears rigid and counter-productive. We encourage evaluators to change set-up or make alterations to the prototype in the middle of a test if they believe it will help them answer important questions about the use of the system. Since TA testing is not a classical laboratory-style scientific testing method evaluators may feel they need to support the formative test results with summative measures. This need for bolstering a usability claim is discussed by (Carter & Yeats, 2005) who points at highlights videos as one way of providing such evidence. Researchers are encouraged to search for other, less expensive methods, for backing up usability results.

Acknowledging the work of Boren and Ramey (2000) this study aims at providing a needed description of how usability evaluation is conducted in practice. Two limitations are worth mentioning. First, we have only collected data in seven companies. Obviously, there are great variations in how usability work is conducted in those companies, which we have not touched upon. A goal for future work should be to collect more coarse-grained data, which would capture the process of usability evaluation in a greater number of companies. Second, we have mainly focused on test sessions. Thus, we did not explore the relation between test sessions and the feedback given to customers; nor did we collect any material on the planning of tests.

Conclusion

We have presented an explorative study of how usability professionals conduct think-aloud tests. It suggests that think-aloud tests might not get sufficiently analyzed. We see a tendency that evaluators end up focusing too much on already known problems, and that the questions they ask during a test seem to concern problems that the user expects, rather than problems actually experienced during the test. The tests were to some extent shaped by practical realities and by some evaluators' adherence to a strict, laboratory-style procedure. Finally evaluators seem to prioritize problems regarding usability over problems regarding utility, when they conduct think-aloud tests.

We encourage further work on methods for fast-paced analysis. Methods and procedures for investigating the utility and probing for users' perception of a system may also be of value for evaluators. Practitioners are advised to more systematically capture and discuss observations from a test. Questions about the practical relevance of the system evaluated could be one way to address utility issues. Investigating problems that are experienced rather than expected may also improve think-aloud tests.

References

- Arnowitz, J., Gray, D., Dorsch, N., Heidelberg, M., & Arent, M. (2005). The Stakeholder Forest: Designing an Expense Application for the Enterprise. *Proceedings of CHI'05*, 941-956.
- Beyer, H., & Holtzblatt, K. (1998). *Contextual Design*. San Francisco, CA, Morgan Kaufman Publishers.
- Boivie, I., Åborg, C., Persson, J., & Löfberg, B. (2003). Why Usability Gets Lost or Usability in in-House Software Development. *Interacting with Computers*, 15, 4.
- Boren, M., & Ramey, J. (2000). Thinking Aloud: Reconciling Theory and Practice. *IEEE Transactions on Professional Communication*, 43, 3, 261-277.
- Carter, L., & Yeats, D. (2005). The Role of Highlights Video in Usability Testing: Rhetorical and Generic Expectations. *Technical Communications*, 52, 2, 1-7.
- Chi, M. (1997). Quantifying Qualitative Analyses of Verbal Data: A Practical Guide. *The Journal of the Learning Sciences*, 6, 3, 271-315.
- Cockton, G., Lavery, D., & Woolrych, A. (2003). Inspection-Based Evaluations. In J. Jacko, & A. Sears, *The Human-Computer Interaction Handbook*, 1118-1138, Lawrence Erlbaum Associates.
- Cockton, G., Woolrych, A., Hall, L., & Hindmarch, M. (2003). Changing Analysts' Tunes: The Surprising Impact of a New Instrument for Usability Inspection Method Assessment. *People and Computers XVII: Proceedings of HCI'03*, 145-162.
- Dumas, J. (2003). User-Based Evaluations. In J. Jacko, & A. Sears, *The Human-Computer Interaction Handbook*, 1093-1117, Lawrence Erlbaum Associates.
- Dumas, J., & Redish, J. (1999). *A practical guide to usability testing*. Oregon, USA, Intellect Books.
- Dumas, J., Molich, R., & Jeffries, R. (2004). Business: Describing usability problems: Are we sending the right message? *Interactions*, 11, 4, 24-29.
- Ericsson, K., & Simon, H. (1993). *Protocol Analysis: Verbal Reports As Data (Revised Edition)*. Cambridge, MA, MIT Press.
- Gulliksen, J., Boivie, I., Persson, J., & Hektor, A. L. (2004). Making a Difference - a Survey of the Usability Profession in Sweden. *Proceedings of Nordichi 2004*, 207-215.
- Hertzum, M. (1999). User Testing in Industry: A Case Study of Laboratory, Workshop, and Field Tests. *Proc. ERCIM Workshop on User Interfaces for All*, 59-72.
- Hertzum, M., & Jacobsen, N. (2001). The Evaluator Effect: A Chilling Fact About Usability Evaluation Methods. *International Journal of Human-Computer Interaction*, 13, 421-443.
- Hornbæk, K., & Frøkjær, E. (2005). Comparing usability problems and redesign proposals as input to practical systems development. *ACM Conference on Human Factors in Computing Systems*, 391-400.
- Hornbæk, K., & Frøkjær, E. (2004). Two Psychology-Based Usability Inspection Techniques Studied in a Diary Experiment. *Proceedings of Nordichi 2004*, 3-12.
- Iivari, N. (2005). Usability Specialists - 'a Mommy Mob', 'Realistic Humanists' or 'Staid Researchers'? An Analysis of Usability Work in Software Product Development. *Proceedings of Interact 2005*, 418-430.
- Jacobsen, N., & John, B. (2000). Two Case Studies in Using Cognitive Walkthroughs for Interface Evaluation. Retrieved from School of Computer Science Technical Report CMU-CS-00-132, Carnegie Mellon University: <http://reports-archive.adm.cs.cmu.edu/anon/2000/CMU-CS-00-132.pdf>

- Jeffries, R., Miller, J., Wharton, C., & Uyeda, K. (1991). User interface evaluation in the real world: A comparison of four techniques. *ACM Conference on Human Factors in Computing Systems*, 119-124.
- John, B. (2004). Beyond the UI: Product, Process and Passion. *Proceedings of Nordichi 2004*, 285-286.
- John, B., & Mashyna, M. (1997). Evaluating a Multimedia Authoring Tool. *Journal of the American Society of Information Science*, 48, 9, 1004-1022.
- John, B., & Packer, H. (1995). Learning and Using the Cognitive Walkthrough Method: a Case Study Approach. *Proceedings of CHI'95*, 429-436.
- Karat, C., Campbell, R., & Fiegel, T. (1992). Comparison of Empirical Testing and Walkthrough Methods in Usability Interface Evaluation. *Proceedings of CHI'92*, 397-404.
- Molich, R. (2003). Discount user testing. Retrieved from www.dialogdesign.dk
- Molich, R., Ede, M., Kaasgaard, K., & Karyukin, B. (2004). Comparative Usability Evaluation. *Behaviour & Information Technology*, 23, 1, 65-74.
- Nielsen, J. (1992). Finding Usability Problems Through Heuristic Evaluation. *Proceedings of CHI'92*, 373-380.
- Nielsen, J. (1993). *Usability Engineering*. San Francisco, CA, Morgan Kaufmann Publishers.
- Pace, S. (2004). Grounded Theory of the Flow Experiences of Web Users. *International Journal of Human-Computer Studies*, 60, 347-363.
- Sawyer, P., Flanders, A., & Wixon, D. (1996). Making a Difference - The Impact of Inspections. *Proceedings of CHI'96*, 376-382.
- Spencer, R. (2000). The Streamlined Cognitive Walkthrough Method, Working Around Social Constraints Encountered in a Software Development Company. *Proceedings of CHI'2000*, 353-359.
- Strauss, A., & Corbin, J. (1998). *Basics of Qualitative Research - Techniques and Procedures for Developing Grounded Theory*. California, Sage Publications.
- Szczur, M. (1994). Usability Testing - on a Budget: a NASA Usability Test Case Study. *Behaviour & Information Technology*, 13, 106-118.
- Vredenburg, K., Mao, J., Smith, P., & Carey, T. (2002). A Survey of User-Centered Design Practice. *Proceedings of CHI'02*, 472-478.
- Wilson, S., Bekker, M., Johnson, P., & Johnson, H. (1997). Helping and Hindering User Involvement - a Tale of Everyday Design. *Proceedings of CHI'97*, 178-185.
- Wixon, D. (2003). Evaluating Usability Methods: Why the Current Literature Fails the Practitioner. *Interactions*, 10, 4, 29-34.
- Zirkler, D., & Ballman, D. (1994). Usability Testing in a Competitive Market: Lessons Learned. *Behaviour and Information Technology*, 13, 1&2, 191-197.

Usability work: A Human Activity²

Mie Nørgaard

Department of Computer Science
University of Copenhagen
mien@diku.dk

Abstract

Much work on usability has a clear human perspective, such as making usability results more useful for developers. Yet, most work end up detaching usability work from human activities in its aim to isolate specific phenomena important to the quality and impact of evaluation results. This paper argues that researchers and practitioners could gain from understanding usability as a human activity involving, for example, learning about and understanding usability issues, and collaborating to improve usability.

Introduction

In the early seventies prominent researchers such as Naur (1971) (for the English translation see (Naur, 1992), and later Boehm (1991) emphasized the importance of the human factor when understanding work on computer systems. These thoughts have also influenced usability work. Still, much of the work on usability concern methods, procedures, or how to report problems in a way that leads to most fixes. Of course, work on how to describe and present results from usability evaluations implicitly concern humans, namely the receivers. Yet, work that aim to isolate important phenomena for example in the use of usability evaluation methods (UEMs), seem often to view usability detached from human activities such as learning and collaborating.

² This paper was originally published for the COST294-MAUSE workshop on downstream utility: the Good, the Bad, and the Utterly Useless Usability Evaluation Feedback, November 6th, 2007, Toulouse, France

To improve the downstream utility of usability work, researchers and practitioners might find it useful to understand usability in terms of activities such as learning about and understanding usability issues, and collaborating with other stakeholders to improve usability. Indeed, borrowing the terms of Naur, we might gain from understanding usability work as a human activity (Naur, 1992), meaning that usability work must not solely be understood as the development and use of certain evaluation methods, but also in terms of creativity, personality, individual professional goals, learning, and collaboration, to mention a few relevant aspects.

Aspects of usability

Literature on usability views usability work and results from a diversity of angles. Traditionally, work on usability has understood the results of usability evaluations as a presentation of problems that is quick and easy to use, see for example (Redish, Bias, Bailey, Molich, Dumas, & Spool, 2002; Dumas & Redish, 1993). More recent, persuasiveness has been mentioned as a key factor for usability's value and impact, and it has been coined by terms such as relevance, salience, reliability and quantity (Law, 2006). Persuasiveness has also been discussed by Nørgaard and Høegh (2008) in terms of argumentation theory. Other work on the presentation of usability results concerns for example the value of redesign proposals (Hornbæk & Frøkjær, 2005), or prioritized results (Hertzum, 2006).

Furniss et al. (2007a) relate themes from Resilience Engineering to usability work, and argue that usability is an activity that needs to be adjustable and flexible, as when usability work adjusts to changing contextual factors. The same view has been touched upon by Nørgaard and Hornbæk (2006), who report that usability experts feel the need to adjust their usability testing to changing conditions. These views relate usability work to the organisation in which it takes place, and suggests that usability is an activity that must be understood together with other activities and contextual factors, and not as a stand-alone activity that can be studied out of context.

The work discussed above, which is mainly related to methods, reporting styles, or to organisational factors, clearly has a human perspective. For example, the attempt to produce redesign proposals that can inspire developers, shows concern for the humans who receive the evaluation results. Still, it does not cover usability as a human activity that is dependent on how humans learn and collaborate.

Let us next investigate how the role of the human is discussed in usability literature.

Usability as a human activity

Dumas (in Redish, Bias, Bailey, Molich, Dumas, & Spool, 2002) has argued that the relationship between developer and usability expert might be the most important factor for usability's success, more important than, for example, how usability results are fed back to developers. Others have made similar observations on the importance of human relationships, such as the relationship between usability expert and customer, users and stakeholders, and so on (Bennet & Karat, 1994; Furniss, Blandford, & Curzon, 2007b; Wixon & Wilson, 1997). In fact, Bennet and Karat (1994) argue that finding ways to facilitate collaboration between stakeholders to usability is a most urgent matter for HCI research.

Along these lines I suggest that researchers and practitioners should understand usability as a human activity rather than as a matter of methods and procedures. Such an understanding should help us focus on human activities such as collaboration and learning between stakeholders. It may also help us see stakeholders not only as professionals but as individuals who work together in professional or cross-professional groups.

The Participatory Design tradition, with its strong focus on collaboration, does perhaps best reflect the understanding of usability as a human activity (Bødker & Buur, 2002). Still, Participatory Design focuses on the beginning of the development process, where much has the form of sketches. It has not had much effect on the further development, such as for example the evaluation of prototypes of full-functioning systems.

I argue that understanding usability as a human activity will help researchers and practitioners bring the focus on for example collaboration and learning beyond the sketching phase of development, and into other parts of development relevant for usability, namely evaluation. In my opinion we need closer attention to stakeholders' job roles and possibly conflicting goals so as to better understand how to support the successful interaction between stakeholders, such as developers, usability experts, and project managers. We also need to understand stakeholders not only as professionals but as individuals as well. Understanding usability as a human activity could help move the attention from the current somewhat inward focus to a broader and more outward one. As an example broadening the focus on the trouble of implementing usability work in organizations, or

the methods used, to include factors such as how stakeholders learn about and understand usability issues, or how they collaborate to improve usability.

Such a broader focus will move usability studies away from lab-style studies that count how many problems a method identifies, or how many of the identified problems a developer plans to fix. Instead, an ethnographical approach is needed to bring qualitative and quantitative descriptions together to describe the forces at work when professionals conduct usability work in the industry. To give due credit, such work seems to be given increased attention, see for example (Cajander, Gulliksen, & Boivie, 2006; Gulliksen, Boivie, Persson, Hektor, & Herluf, 2004; Uldall-Espersen & Frøkjær, 2007). Still, we need more field work that studies for example how usability work is organised and conducted in different and diverse organisational settings. We need to understand how ambitions and goals for the usability of products come to be, and how inter-personal and inter-professional relationships reflect on usability work in an organisation. Similarly, knowing more about precisely how great designs or design improvements came to be, would also be valuable to usability researchers and practitioners.

To answer all these questions (and many more) researchers need to move out of the offices, and into the field.

To conclude, usability is about UEMs, redesigns, problem severity and many other issues related to methods and results. Yet, I argue that usability work is mainly about human activities such as learning and collaboration. As a consequence improving usability is not mainly about getting more resources to test a system more often, or to study systems in more detail. Improving usability is about understanding usability work as a diverse collaborative learning process, where both professional and personal relations between stakeholders are crucial. To better understand the complex nature of usability work, and how we might improve downstream utility, researchers need to go where the action is—namely the industry—to do more studies.

References

- Bennet, J., & Karat, J. (1994). Facilitating effective HCI design meetings. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Celebrating Interdependence*, 198-204.
- Boehm, B. W. (1991). *Software Risk Management: Principles and Practices*. IEEE Software, 8, 1.
- Bødker, S., & Buur, J. (2002). The design collaboratorium - a place for usability design. *ACM Transactions on Computer-Human Interaction (TOCHI)*.
- Cajander, Å., Gulliksen, J., & Boivie, I. (2006). *Management Perspectives on Usability in a Public Authority - A Case Study*. *Proceedings of NordiCHI 2006*.
- Dumas, J., & Redish, J. (1993). *A practical guide to usability testing*. Oregon, USA, Intellect Books.
- Furniss, D., Blandford, A., & Curzon, P. (2007a) *Resilience in Usability Consultancy Practice: the Case for a Positive Resonance Mode*. *Workshop on Resilience Engineering*, 25-27 June, Vadstena, Sweden.
- Furniss, D., Blandford, A., & Curzon, P. (2007b). *Usability Work in Professional Website Design: Insights From Practitioners' Perspectives*. In E. Law, E. Hvannberg, & G. Cockton, *Maturing Usability: Quality in Software, Interaction and Value*, 144-167, Springer, London.
- Gulliksen, J., Boivie, I., Persson, J., Hektor, A., & Herluf, L. (2004). *Making a Difference - a Survey of the Usability Profession in Sweden*. *Proceedings of Nordichi 2004*, 207-215.
- Hertzum, M. (2006). *Problem Prioritization in Usability Evaluation: From Severity Assessments to Impact on Design*. *International Journal of Human-Computer Interaction*, 21, 2, 125-146.
- Hornbæk, K., & Frøkjær, E. (2005). *Comparing usability problems and redesign proposals as input to practical systems development*. *ACM Conference on Human Factors in Computing Systems*, 391-400.
- Law, E. (2006). *Evaluating the downstream utility of user tests and examining the developer effect: A case study*. *International Journal of Human-Computer Interaction*, 21, 2, 147-172.
- Naur, P. (1992). *Problem Formulation - The Fertile Soil of the Software Development Project*. In P. Naur, *Computing: A Human Activity*. New York, ACM Press/Addison-Wesley.
- Naur, P. (1971). *Problemformulering - Edb-Projektets Grobund*. Data, 1.
- Nørgaard, M., & Hornbæk, K. (June 26th-28th, 2006). *What Do Usability Evaluators Do in Practice? An Explorative Study of Think-Aloud Testing*. *Proceedings on the 6th ACM Conference on Designing Interactive Systems (DIS 2006)*. Penn State, Pennsylvania.
- Nørgaard, M., & Høegh, R. T. (2008, February 25th-27th, Cape Town, South Africa). *Evaluating Usability - Using Rhetorical Models to Improve the Persuasiveness of Usability Feedback*. *Proceedings of the 7th ACM Conference on Designing Interactive*

Systems (DIS2008).

Redish, J., Bias, R., Bailey, R., Molich, R., Dumas, R., & Spool, J. (2002). Usability in practice: Formative usability evaluations - Evolution and revolution. ACM Conference on Human Factors in Computing System, Minneapolis, Minnesota, 885-890.

Uldall-Espersen, T., & Frøkjær, E. (2007). Usability

and software development: Roles of the stakeholders. Proceedings of HCI2007, July 22.-27., Beijing, China, 642-651.

Wixon, D., & Wilson, C. (1997). The usability engineering framework for product design and evaluation. In M. Helander, T. Landauer, & P. P., Handbook of Human Computer Interaction, 653-688. North-Holland, Elsevier Science.

User testing in the combat zone³

Mie Nørgaard

University of Copenhagen
Universitetsparken 1
DK- 2100 Copenhagen
phone +45 3532 1446
mien@diku.dk

Janus Rau

IO Interactive
Kalvebod Brygge 35-37
DK- 1560 Copenhagen
phone +45 2538 0061
januss@ioi.dk

Abstract

This paper describes the how IO Interactive, a producer of computer games such as the Hitman series, has taken the first step towards working with usability evaluations in a structured manner. The paper describes the usability team's first experiences with testing computer games and their work to integrate usability evaluation in the design of computer games. Finally, the paper identifies five categories of challenges that are vital for the usability team's success; justifying the costs of usability evaluation towards management; identifying structured work procedures that leaves room and opportunity for usability evaluation; identification and use of new methods to support the study of game-specific issues such as replayability and game play; the ability to make alliances with important colleagues and managers; and identifying the people responsible for fixing usability issues.

Introduction

Computer games are on a difficult mission. On one hand they must be accessible and intuitive to use, on the other they must avoid being so easy to use that they become boring. Thus a computer game's value rests heavily on its ability to present the user with exactly the right amount of challenge so that the game is easy to learn but difficult to master. Computer games focus on activities and rules,

³This paper was originally published for Workshop on Methods for Evaluating Games - How to measure Usability and User Experience in Games, The International Conference on Advances in Computer Entertainment Technology (ACE'07), June 13-15, 2007, Salzburg, Austria.

and are designed with chance elements (Wixon, 2006). In this respect computer games differ substantially from office systems, which primary goals are fast, easy and efficient interaction, and which focus on results. Accordingly, usability evaluation methods (UEMs), which are developed to test office systems, often perform poorly when applied to computer games because they do not take system specific issues such as game play and re-playability into account.

This paper sets out to describe some of the experiences and difficulties the computer games company IO Interactive (IOI) had during its first attempt to use usability evaluation as a tool in the development process. The paper is based on an interview and discussions with QA-Manager Janus Rau. The semi-structured interview was conducted and transcribed in December 2006 as part of a recent research project on usability challenges in different industries.

Related work

Over the last 20 years much work has been done on developing and describing various usability evaluation methods. The think aloud protocol (Ericsson K. H., 1993; Rubin, 1994) has been received with open arms by an industry that is increasingly aware that usability matters. The method is in fact so popular that people have named it 'the golden standard'. The development of expert methods such as heuristic evaluation (Nielsen, 1992; Nielsen, 1993) and cognitive walkthrough (Wharton, Rieman, Lewis, & Polson, 1994) has boosted the usability field by providing a quick and low-cost alternative or supplement to think aloud testing. Further, much work has been done to evaluate and compare UEMs in order to improve on methods and procedures, see for example (Cockton, Lavery, & Woolrych, 2003; Dumans, 2003; Hornbæk & Frøkjær, 2005; Jeffries, Miller, Wharton, & Uyeda, 1991; John & Mashyna, 1997; Karat, Campbell, & Fiegel, 1992; Nielsen, 1992).

Simultaneously with the upcoming of usability, the popularity of computer games has boomed, and games are today a billion-dollar industry. Some work has dealt with how usability professionals use the think aloud protocol to evaluate office and web-based systems (Nørgaard & Hornbæk, 2006), but records on how usability in computer games is evaluated are scarce. In his overview article Helms Jørgensen explains the lack of descriptions of evaluation praxis with 'Microsoft [being] the only example of major game developers having seriously taken up usability approaches' (Jørgensen, 2004).

Though many of the existing UEMs can be used to test at least aspects of a computer game (Wixon, 2006), games are substantially different from office systems and the usability of computer games need to be tested on its own terms. Several steps have been taken to facilitate the evaluation of usability in games during the years. In 1982 Malone constructed a list of heuristics for instructional games (Malone, 1982) and the RITE method emphasized the value of rapid changes and close involvement of decision makers such as program managers or game designers (Medlock, Wixon, Terrano, Romero, & Fulton, 2002). Fabricatore and Rosas (2002) later created an empirically based model with prescriptions and recommendations on how to design games, and Desurvire and Toth. (2004) more recently developed a set of heuristics for playability (HEP) based on current literature and expert reviews. The indisputable value of this and related work aside, usability evaluation is still not necessarily a well-integrated part of the development of computer games, and usability practitioners are still in want for better methods and procedures to help them work specifically with the improvement of usability in games.

Since we believe that current challenges and methodological shortcomings must drive the improvement of methods and procedures, we next present a case description of IOI's experiences with integrating usability evaluation in their development process.

The IOI case

The business

IO Interactive is a Danish producer of computer games, and has since 1998 produced successful games such as the Hitman games and Freedom Fighters. IOI develops, designs and produces interactive entertainment for the major platforms on the global market, and leaves marketing, sales and distribution to its owners and publishers, the SCi/ Eidos group.

According to the company's website both the Hitman games and Freedom Fighters have received numerous awards and nominations as recognition of their quality over the years, such as the BAFTA award and Gamespot's 'Best of' award. The reviews from several large games magazines such as PS2 Magazine, IGN and GamePro also conclude that IOI produces highly successful games with strong appeal to the gaming audience. However, though the Hitman series are highly recognized for its game play, these games are also widely known as being difficult to access for novices, according

to QA-Manager Janus Rau. And since the lack of usability is likely to have kept many users from getting value for their money or even purchasing the product in the first place, IOI has recently increased its attention on evaluating usability. Accordingly, between the last release; *Hitman: Blood Money* and this year's release; *Kane & Lynch: Dead Men* (set to autumn 2007) an important step has been taken to address the usability of new games: including usability evaluation in the development process. In the following we describe the work leading up to this point.

Usability praxis in IOI

The computer games industry has a lot of focus on its users, but not on involving users in the design process. 'Even though the game designers are very knowledgeable about game design theory, and work very hard on making the game accessible, the real users were earlier practically kept out of the production process', explains Janus Rau, 'however the risk entailed in this approach is, that you can end up designing with a bias towards own tastes and preferences'. Such tendencies are not unique; Ernest Adams, a former game designer for EA and co-founder of IGDA, describes a similar scenario; 'I've been working for a major game developing company for 8 years and I've never seen a methodologically sound study of who the players are - game design is based on common wisdom and guesses - designers build games for themselves' [Jørgensen, 2004].

Convincing the management to test

IOI does not have a full time usability team, and thus any activities are highly dependent on enthusiastic employees. The usability work is mainly conducted by the QA-Manager, as a side interest, and a part time usability assistant. Since the work on *Hitman: Blood Money* this team has convinced a sometimes sceptic management that usability evaluations are needed if IOI wants to address their users' demands, and stand a better chance delivering well-designed games on time. In order to convince the management Rau and colleague criticized the at-the-time current procedures for usability evaluation; a few weeks before the 1st submission deadline the mother company would carry out a traditional user test, but the poor timing meant that the results from the tests were either rejected as being unrealistic within the given time frame or only reluctantly accepted as valid. Developers and game designers at IOI thus experienced that usability evaluation did little more than point fingers at a development team that had no time to fix the identified usability issues.

IOI's usability team used this inadequate work process as an argument to persuade management about the value of IOI conducting their own

usability evaluations during development. The outlook to optimize game development made an impact on the management, who previously had been sceptic about the relevance of usability evaluation. 'I have explained that the purpose is not to ask the users what they personally like, or what colours they prefer, but rather to test how they actually use the game and then utilize this information in the development process', Rau explains, 'but those arguments did not always seem to sink in.'

Methods and procedures used for testing

The methods used for usability evaluation in IOI have varied depending on the part of the game being evaluated. A menu system can fairly easily be evaluated using the think aloud protocol, but the need to study the use of the actual game calls for other methods. 'While a user plays the computer game, so much happens that it is not possible to also think aloud', Rau explains, 'I tried using the think aloud protocol once, but it did not work. Users simply kept dying, which rendered the tests completely useless'.

To test the interaction between player and game the evaluators at IOI observe users play the game, video record the séance, and analyze it afterwards. Developers are invited to the test either as silent observers, or as more active participants, interacting with the users. 'I aim to involve users in the development but also to involve developers in the use process, because the developers are more likely to acknowledge the relevance of the usability work if they feel they have a stake in it', Rau explains. For the same reason developers contribute to the development of test tasks, and can ask for certain areas to be examined more thoroughly. The play session thus functions as a contextual interview, where the evaluator sits by the user's side, observes the interaction, and asks elaborating questions.

The analysis of a test ends up identifying around 30 problems. Rau and his colleague then select the seven or eight most important ones, which they know can be fixed within deadline. Since the goals of the game play are never meticulously identified and described by the game designers, the evaluators need to have a detailed idea of the game designers' visions for the game before they can estimate what is a usability problem, and what is an intended challenge for the player. In order to estimate which problems are fixable within the given deadline the evaluator also has to have detailed knowledge of how the system is build, and is thus dependent on being in a continuous dialogue with the game designers.

50 users have until now participated in tests on the upcoming release, and recommendations

from the tests have been received with great interest by game designers.

Challenges and shortcomings

The usability challenges for IOI can be placed in five categories; justifying the costs, work procedures, user involvement, collaboration and alliances, and responsibility.

Next, we describe and discuss the categories in turn, briefly suggesting possible solutions.

Justifying the costs

One of the key challenges for IOI's usability team is to persuade the management to allocate time and money to evaluate an upcoming product's usability. To do this the team needs to justify adding what seems to be yet another time consuming element to the game design process. Management may truthfully argue that games have done fine without much usability evaluation so far, and may question whether usability studies will actually return the investment. However IOI has up until now designed games to users who are very similar to the people developing the games. This may be one of the reasons why games such as the Hitman series have had success despite the lack of user studies. Nevertheless, if the company is to continue its growth in the future it needs to expand its target audience to include other types of users, and this is likely to be more successful if user studies are prioritized. A quick look at IOI's most recent portfolio actually reveals that the games are already becoming more aimed for the mass market, and suggests that thorough user studies are needed in the near future.

Rau mentions that game producers can be particularly difficult to persuade since they are often concerned with short-term goals like making the next deadline and may not respond to arguments for usability's long-term benefits. 'I do not consider this a big problem though, only a constant reminder that we have to be able to argue the case of usability evaluation', he adds. Thus convincing arguments about both short- and long-term costs and benefits of usability studies are needed if management is to be persuaded to allocate means to the usability team.

Work procedures

Despite the company's increased focus on usability, IOI has no structured process for evaluating a product's usability; work is timed and conducted ad hoc to suit the development process. The usability team calls for not only a structured process for conducting tests at crucial points in the development process, but also for a process that describes how the results are handled after a test in order to secure that the results actually have an

impact on the product.

Though a growing number of colleagues are interested in usability work, Rau finds it challenging to actually involve them in tests and results. Most of his colleagues are on a tight schedule, and since usability evaluation is not a mandatory part of game development in IOI as it is, Rau has difficulty convincing colleagues that they need to spend valuable time discussing usability issues. He suggests further that the very act of involving users in the design of games may even leave some colleagues to feel threatened.

On a more practical note, the time and effort used by the development team to make a stable version of the game to test, is not a part of the production schedule as it is.

Generally a more thoroughly defined process that identifies when and how usability studies are used in game development will prove helpful to the usability team. Similarly, agreeing on who is responsible for and has the power to make decisions about usability priorities will also help strengthening usability's impact on games development in IO Interactive as well as attempts to demystify user involvement. Finally, shaping the development process to support usability evaluation, like making room for creating and testing early prototypes, may also prove helpful.

User involvement

While the think aloud protocol may be useful to test menus or how to set-up the game, other methods are needed to evaluate areas such as player experience and deep game play. Rau explains that using the think aloud protocol to get information about what users find difficult when playing is impossible; in an action game it is simply too confusing for the users to talk and play at the same time. At the moment the usability team studies how users play a game by video recording the interaction, and analyzing it after the test session. This approach is useful to unveil problems such as having trouble understanding or navigating the game. Still, a broader palette of methods and procedures are called for to evaluate issues such as game play, degree of challenge, social usability and re-playability. While traditional usability evaluation methods see computer games as software, complementary methods that also see the computer game as a game are needed (Barr, Noble, & Biddle, 2007).

Collaboration and alliances

Since the game designers' visions are rarely entirely documented, and usability evaluators are dependent on knowing these visions to understand what is a usability issue and what is an intended challenge, the collaboration and communication

between game designers and usability evaluators are crucial. IOI might consider strengthening this collaboration by defining formats or procedures through which game designers and usability evaluators can share visions that are relevant to usability. This might facilitate game designers and evaluators agreeing about which part of the game is intended to be challenging (for instance a part of the story line that is not to be understood until a certain point in the game) and which parts are not intended to be a challenge (for instance how to save a game).

Another challenge for the usability team in IOI is to get the relevance of its work acknowledged by the colleagues. An important competence for a usability evaluator is thus the ability to work strategically. Rau and his colleague have a standing agreement with the game designers that testing with users should not result in a long list of wishes for new features, since deadlines are hard to meet as it is. In order to strengthen usability's impact Rau and his colleague also seek to make alliances with those managers who take an interest in usability. 'There is one member of the game management who is really into usability, and welcomes all the input I have', says Rau. Thus, knowing who and how to influence colleagues and managers in order to push the usability work further is a vital criterion for success.

Responsibility

Once tests are analyzed and recommendations are available, the usability team faces a new challenge; to whom shall the usability feedback be directed, and who is responsible for carrying out which usability recommendations? Rau explains how different members of the game management are responsible for different parts of the game, but that some usability results simply fall between areas of competence. And if convincing a member of the game management to deal with usability issues in his own domain is difficult, convincing him to deal with issues outside of his domain is practically impossible. One important challenge for IOI is thus to clearly identify areas of responsibility and facilitate a forum where usability issues that do fall between areas of competence are discussed and handled instead of put on hold or dropped on the floor.

We have presented some of the key challenges for IO Interactive's usability team and its attempts to integrate usability work in the development of computer games. Some of these challenges such as justifying the costs are common for any software company in the process of maturing its view on usability. Some, such as the need for new UEMs, are very specific for the area of game design. Nonetheless, dealing with all of these chal-

lenges is of vital importance if computer game companies are to take user studies and computer games to the next level.

References

- Barr, P., Noble, J., & Biddle, R. (2007). Video Game Values: Human-Computer Interaction and Games. *Interacting with Computers*, 19, 180-195.
- Cockton, G., Lavery, D., & Woolrych, A. (2003). Inspection-Based Evaluations. In J. Jacko, & A. Sears, *The Human-Computer Interaction Handbook*, 1118-1138, Lawrence Erlbaum Associates.
- Desurvire, H. M., & Toth, J. (2004). Using heuristics to evaluate the playability of games. *CHI '04 extended abstracts on Human factors in computing systems*, 1509-1512 .
- Dumas, J. (2003). User-Based Evaluations. In J. Jacko, & A. Sears, *The Human-Computer Interaction Handbook*, 1093-1117, Lawrence Erlbaum Associates.
- Ericsson, K. H. (1993). *Protocol Analysis: Verbal Reports As Data (Revised Edition)*. Cambridge, MA, MIT Press.
- Fabricatore, C. M., & Rosas, R. (2002). Playability in Action Video Games: A Qualitative Design Model. *Human Computer Interaction*, 17, 4, 311-368.
- Hornbæk, K., & Frøjkær, E. (2005). Comparing usability problems and redesign proposals as input to practical systems development. *ACM Conference on Human Factors in Computing Systems*, 391-400.
- Jeffries, R., Miller, J., Wharton, C., & Uyeda, K. (1991). User interface evaluation in the real world: A comparison of four techniques. *ACM Conference on Human Factors in Computing Systems*, 119-124.
- John, B., & Mashyna, M. (1997). Evaluating a Multimedia Authoring Tool. *Journal of the American Society of Information Science*, 48, 9, 1004-1022.
- Jørgensen, A. (2004). Marrying HCI/Usability and Computer Games: A Preliminary Look. *Proceedings of NordiChi '04, October 23.-27. 2004, Tampere, Finland*, 393-396.
- Karat, C., Campbell, R., & Fiegel, T. (1992). Comparison of Empirical Testing and Walkthrough Methods in Usability Interface Evaluation. *Proceedings of CHI'92*, 397-404.
- Malone, T. (1982). Heuristics for Designing Enjoyable User Interfaces: Lessons From Computer Games. *Proceedings of HumanFactors in Computer Systems, Gaithersburg, Maryland*, 63-68.

- Medlock, M., Wixon, D., Terrano, M., Romero, R., & Fulton, B. (2002). Using the RITE Method to Improve Products; a Definition and a Case Study. Proceedings of Usability Professionals Association (UPA), Orlando, FL .
- Nielsen, J. (1992). Finding Usability Problems Through Heuristic Evaluation. Proceedings of CHI'92, 373-380.
- Nielsen, J. (1993). Heuristic evaluation. In J. Nielsen, & R. Mack, Usability inspection methods, 25-62, John Wiley & Sons.
- Nørgaard, M., & Hornbæk, K. (2006). What Do Usability Evaluators Do in Practice? An Explorative Study of Think-Aloud Testing. ACM Conference on Designing Interactive Systems (DIS 2006). Penn State, Pennsylvania.
- Rubin, J. (1994). Handbook of Usability Testing: How to Plan, Design and Conduct Effective Tests. New York, John Wiley & Sons inc.
- Wharton, C., Rieman, J., Lewis, C., & Polson, P. (1994). The Cognitive Walkthrough Method: A Practitioner's Guide. In J. Nielsen, & R. Mack, Usability Inspection Methods, 105-141, John Wiley & Sons inc.
- Wixon, D. (2006). What works? Interactions, 13, 4, 18-19.

Working Together to Improve Usability: Challenges and Best Practices⁴

Mie Nørgaard

University of Copenhagen
mien@diku.dk

Kasper Hornbæk

University of Copenhagen
kash@diku.dk

Abstract

In theory, usability work is an important and well-integrated activity in developing software. In practice, however, collaboration on improving usability is ridden with challenges relating to conflicting professional goals, tight project schedules, and unclear usability findings. We study those challenges through 16 interviews with software developers, usability experts, and project managers. Four key challenges to successful interaction between stakeholders are identified: poor timing when delivering usability results, results lacking relevance, little respect for other disciplines, and difficulties sharing important information. We discuss practices that address these challenges, and present four guidelines to support the collaboration and professional relationship among developers, usability experts, and project managers. Our observations are further discussed as encompassing multiple perspectives and as a collaborative cross-professional learning process.

This paper was originally published in 2008 as a Technical report from Copenhagen University Dept. of Computer Science, <http://www.diku.dk/publikationer/tekniske.rapporter/rapporter/08-01.pdf>

Introduction

Through their work, usability professionals aim to improve the usability of computer systems. To do

this, they seek to inform and influence design decisions, for instance by conducting usability evaluations of systems, by instigating design changes through persuasive reports, and by strengthening the collaboration with colleagues who also have a stake in designing and implementing the systems.

Accordingly, increasing the impact of usability work on system design and implementation can be approached in several ways. Such ways include attempts to improve the quality of usability evaluation methods by trying to identify which method works best in certain contexts (Karat, Campbell, & Fiegel, 1992), empirically describing strengths and shortcomings of a particular usability evaluation method, recommending ways of combining methods (Uldall-Espersen, Frøkjær, & Hornbæk, 2007), or investigating how to present the results of evaluations so as to facilitate changes in the design (Hvannberg, Law, & Larusdottir, 2007). Because usability is closely related to the work of for example project managers and developers, one may also seek to improve the collaboration between usability experts and other stakeholders (Bødker & Buur, 2002; Gulliksen, Boivie, & Göransson, 2006).

The motivation for this paper is that while the literature is strong on most points above, little research concerns the last point, in particular the practical challenges of how to collaborate to improve usability. We seek to strengthen the literature by investigating real-world collaboration on usability-related issues across a range of organizations. To do so, we conduct a grounded theory analysis of 16 interviews with 20 stakeholders, and, based on the perspective of the participants, we seek to answer the following questions:

- a) What do key stakeholders—developers, usability experts, and project managers—consider their main challenges when they cooperate on improving usability?
- b) Which best practices do stakeholders follow to address these challenges to usability work?

The answers to these questions may improve the impact of usability work, for instance by suggesting how to conduct usability work that lessens challenges amongst stakeholders. In relation to research in usability evaluation, the study identifies questions and best practices that we argue deserve the attention of researchers. Our study also extends the existing literature by highlighting the interplay among stakeholders and by analysing not only challenges, but also best practices.

Related work

Part of the literature on strengthening the impact of usability work focuses on usability evaluation methods (UEMs) (Chattrachart & Brodie, 2004; Hertzum & Jacobsen, 2001; Hvannberg, Law, & Larusdottir, 2007; Law & Hvannberg, 2004) or on how evaluation results are reported (American National Standards Institute, 2001; Cockton, Woolrych, & Hindmarch, 2004; Dumas & Redish, 1999; Mills, 1987; Redish, Bias, Bailey, Molich, Dumas, & Spool, 2002; Rubin, 1994). Other contributions look into the context of usability work (Gulliksen, Boivie, & Göransson, 2006; Gulliksen, Boivie, Persson, & Hektor, 2004; Iivari, 2006; Uldall-Espersen & Frøkjær, 2007) or relate the collaboration and communication among stakeholders to the development process (Bennet & Karat, 1994; Bødker & Buur, 2002; Bødker & Krogh, 2001; Hornbæk & Frøkjær, 2005; Madsen & Petersen, 1999; Uldall-Espersen & Frøkjær, 2007). This paper follows the latter trail and views usability work primarily as an organisational activity, in particular the collaboration between three key job roles, cf. Figure 1.

Gulliksen et al. (2006) investigated the work context for usability professionals and suggested that the impact of usability work does not solely depend on usability evaluation methods, but also on support from project management and involvement of stakeholders. Most frequently, involvement of stakeholders in systems development has meant user involvement. For many years user involvement has attracted attention as a means for improving the quality of systems (Boland, 1978; Ives & Olson, 1984; King & Rodriguez, 1981; Robey & Farrow, 1982). As an example, work on participatory design discusses how to strengthen HCI work by involving users in the design process, see for example (Greenbaum & Kyng, 1991; Ehn & Sjögren, 1991; Ehn, 1992). In contrast, the idea of involving other stakeholders, such as developers or project managers in usability evaluation has received less attention. In fact, stakeholder involvement in usability work has mainly been limited to letting developers watch users interact with the system, see for example (Coble, Karat, & Kahn, 1997; Dumas, 1989; Kennedy, 1989; Mills, 1987; Nayak, Mrazek, & Smith, 1995; Redish, Bias, Bailey, Molich, Dumas, & Spool, 2002; Schell, 1986).

Practical insights and case stories, such as presented in (Johnson & Johnson, 1990; La Fasto & Larson, 2002; Winer & Ray, 1994), improve our understanding of how stakeholders collaborate and communicate to improve usability of systems is. For instance, Bennet and Karat (1994) described experiences with using collaborative design meetings to support collaboration and com-

Job descriptions

- Project manager**
- >Serves as contact to management
 - >Serves as contact to customer
 - >Plans workflow
 - >Prioritizes work
 - >Coordinates tasks and people

- Usability expert**
- >Plans and conducts usability tests
 - >Analyses test results
 - >Produces and presents feedback from tests

- Developer**
- >Analyzes and designs solutions
 - >Implements systems
 - >Writes and changes code
 - >Maintains code
 - >Fixes bugs

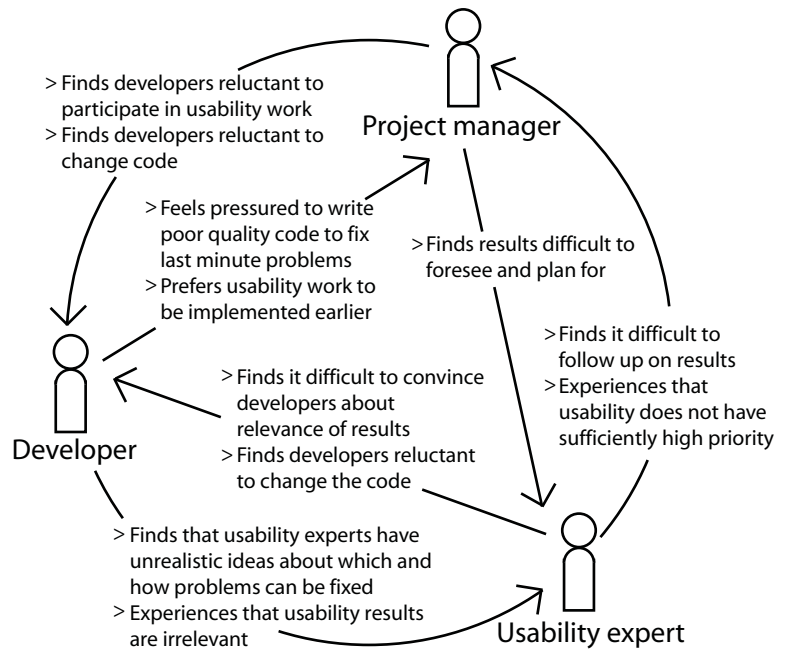


Figure 1: To the left the main activities for a typical developer, usability expert and project manager are described. To the right, the model shows the challenges that these stakeholders face when working together.

munication in HCI. However, they also pointed to major discrepancies between good intentions for effective team work and what is actually practised in the industry. They also identified a gap between intentions about interdisciplinary collaboration and actual work practices as a key challenge for HCI (Bennet & Karat, 1994).

Following the thoughts of Bennet and Karat, we hypothesize that the impact of usability work can be improved by understanding successful usability work as a collaborative process involving different stakeholders such as developers, usability experts, and project managers. This study explores how stakeholders work with and use results from usability evaluations. It does so to identify issues among different groups of professionals, here called cross-professional relationships that may impede usability and evaluation work. The choice of focus does not mean that we do not recognise that other types of work such as studies of user experience or collective design efforts can influence the design and usability of a product. Also, we recognise that the quality of usability evaluation methods, the skill with which they are used, and the format in which results of evaluations are reported to stakeholders are also determinants of how well usability work impacts the development process. We find that the focus on cross-professional relationships is relevant to understanding the context in which evaluation results are made and used by stakeholders who are both professionals and individuals.

Method

Our approach to addressing the two research questions is to conduct and analyze interviews to examine key stakeholders' views on usability work. We use interviews because most work on usability professionals is based on questionnaires (Rosenbaum, Rohn, & Humburg, 2000; Gulliksen, Boivie, Persson, & Hektor, 2004), but see (Iivari, 2006; Gulliksen, Boivie, & Göransson, 2006) for exceptions. Interviews should further allow for richer descriptions of challenges and best practices. We choose stakeholders working as developers, usability experts, and project managers from a variety of different companies to get a richer understanding of usability work. Also, existing literature on usability work predominantly concerns the perspectives of the user (Bødker & Buur, 2002) and the usability professional (Gulliksen, Boivie, Persson, & Hektor, 2004) it rarely concerns developers or project managers, except as described through the perspective of the usability professional. As our aim is to understand a set of work situations and not to test specific theories or hypotheses, we base our methodological approach on grounded theory (Strauss & Corbin, 1998).

Participants

We conducted a total of 16 interviews, each lasting about 1½ hours, with 20 people from the Danish industry, cf. Table 1. Five participants were iden-

Company	Employees		Participants			Company's organizational relation to the usability expert	Type of system
	Denmark	International	D	U	P		
1	800	0	1	1		In house	Banking
2	150	500	1			In house	Games
3	300	0		2		In house	Learning
4	40	0	1	2		In house	e-Government
5	16	0		1		In house and consultancy	External customers/ own development
6	8,500	0	1			In house and consultancy	Booking
7	120	0		1		Consultancy	Homepage and ERP system
8	5	0		2		Consultancy	External customers
9	2,800	0			1	In house	e-Government
10	350	61,000	2			In house	Off-the-shelf and tailored systems
11	8	0		1		Consultancy	External customers
12	220	250	1			In house	Security
13	1,500	59,000			1	In house	Mobile interfaces
14	350	15,000	1			In house	e-Government and off-the-shelf

Table 1: An overview of the participants in the study including data on companies, products and job roles. Four of the 16 interviews included two participants fulfilling the same role. These are marked with the number two in the participants' column. The letters D, U and P in the participants' column refer to: Developer, Usability expert and Project manager.

tified amongst members of a Danish HCI Special Interest Group, the rest were recommended by other participants. Participants had between two and 20 years of professional experience from their current or similar jobs. They comprised 9 usability practitioners, who conduct usability tests and feed the results into the development process, 6 developers, who develop systems and use usability feedback on these, and 5 project managers, who manage system development projects and use usability feedback on systems as part of their job. However, for some participants job roles were not that uniform. Some project managers, for example, had a background in development and some developers also conducted usability work. When referring to participants' job roles, we refer to the predominant job role (see Table 1).

Data Collection and Analysis

Data collection and analysis were done in two phases: (a) an exploratory phase with eight interviews and (b) a focusing phase with eight interviews. In each phase, collection and analyses were interwoven. This was done to explore multiple viewpoints on challenges and best practices, and to develop and follow up on these in subsequent interviews. Next, we explain the two phases.

In phase (a), eight semi-structured interviews were conducted to investigate the issues of work challenges and best practices. To better understand which parts of their jobs participants found challenging, we asked them to describe and exemplify what they found to be particularly difficult in their work. To better understand which tools or techniques participants used to address such work challenges, we prompted for elaborate examples of successful work procedures, events, or techniques they had used or experienced.

The eight interviews were audio-recorded and transcribed. The interviews were compared in order to categorize findings. Eleven categories, covering topics such as methods, job experience, view on usability, and work challenges were identified this way. Each category was further divided into sub-categories by repeating the coding procedure. Finally, the category 'work challenges' was identified as the core category. Work challenges covered specific challenges as well as the best practices that were used to address them. To get finer-grained data about work challenges, the sub-categories were investigated further in eight subsequent interviews (phase (b)). These steps correspond with grounded theory's terms: open coding, axial coding and selective coding (Strauss

& Corbin, 1998).

In phase (b), we transcribed the last eight interviews, and coded these according to the eleven sub-categories. Coded segments would contain issues such as a description of a work procedure, a comment on a certain type of challenge, or a reference to techniques used to facilitate cooperation in a team. This procedure also builds on grounded theory (Strauss & Corbin, 1998) and follows Chi’s proposal for how to analyze verbal protocols (Chi, 1997). Accordingly, the coding scheme is not developed prior to the conduction of the interviews but after; already conducted interviews serve as inspiration and input to subsequent interviews. When referring to statements or quotes from interviews we refer to the number of the company followed by an D/U/P depending on the interviewee’s job role, for example [1P] for the interview with the project manager from company 1, cf. Table 1.

Findings

In the following we describe four key challenges that complicate work relations among developers, usability experts, and project managers. Important challenges for the cross-professional relationship concern poor timing of usability work, usability results lacking relevance, colleagues showing disrespect for others professional goals, and difficulties related to sharing and getting relevant information, cf. Table 2. While these are not the only challenges they are the most frequent and severe. We present these challenges as aspects of the relationship between two job roles, cf. Figure 1. Then, we present best practices that address key challenges.

The Developer- Usability Expert Relationship

From the developers’ point of view

Four developers report that feedback from usability studies is often useless due to bad timing. The confrontation with problems they do not have time to fix only discourages developers who respond with hostility toward usability. One developer wonders about the usability experts’

feedback practice:

Why don’t they just stop giving feedback when the software has been made (...) It is like if you are building a house and someone suddenly says: “Sorry, I would like you to put in a basement also”. Well, are we supposed to tear the whole house down then? Close to a deadline developers do not have time to do anything but move a few things around. And it is not responsible to change software 14 days before release, anyway [4D].

Four of the developers criticise the results of usability work for often being irrelevant since they do not consider, for instance, how the system is built or how products are sold. To exemplify, one developer explains: ‘Every time he [the usability expert] presented a nice suggestion, we could tear it apart because it simply could not work technically. Not because of the system, but because of how our product is sold’ [6D]. Another developer elaborates on the issue of relevant feedback:

When someone has created a piece of software then he needs intelligent feedback and not: “I don’t really know what the system is doing”. Developers usually take the time to learn how things work, and it is hard to respect people who don’t bother. [4D]

Three developers [4D, 12D, 14D] report that having colleagues who do not fully understand how they work, or what are important professional goals are for a developer, is a major challenge for working with usability. They describe how usability experts hold unrealistic ideas about what

Challenge	Examples	N		
		D	U	P
Timing	Poor timing of usability work Pressure to cut corners	4	4	3
Relevance	Feedback from tests lacks relevance	5	6	5
Respect	Low professional ethos Disrespect for other’s job roles and professional goals	3	6	4
Communication	Difficulty communicating usability results or understanding the domain	5	3	3

Table 2: The four challenges described in this paper. Each subcategory concern both challenges and best practices related to the main theme. The N-column refers to the number of interviews in which a sub-category was found. The letters D, U, and P describe interviews with developers (D), usability experts (U), and project managers (P).

developers can change within a system at a certain point during the development. For instance, some usability requirements cannot be fulfilled because they conflict with the choice of platform or because they interfere with other design decisions.

The data suggest that sometimes developers' reluctance to accept usability results spring from their view of how usability studies are conducted and results are communicated. One developer comments on receiving usability results:

Even though they are not supposed to be a critique of the development work, you tend to defend the choices you have made (...) Especially if they have used some sort of heuristic hocus pocus—then they might point out problems where the developers respond: "But that is just your personal preference" (...) And then getting a report on 70 pages and 417 problems, while you are already thinking about the next steps of the project because the project manager is on your back—well, it is just not exactly what you need (...) I cannot find the time to read 70 pages. [12D]

More than half of the participants (four usability experts, four developers and three project managers) criticize written reports for being useless because they are too long.

From the usability experts' point of view

All six interviews with usability experts show that usability experts are particularly concerned about the persuasive power of feedback. They describe how convincing their audience about the relevance and existence of usability problems can be a difficult task. Not only are some problems difficult to explain in a clear manner, but all usability experts also experience how some usability issues are questioned or dismissed by developers. Usability experts also find developers reluctant to change the system's code, a point confirmed by some of the developers. As an example, one usability expert explains: 'It is a problem to convince developers about the relevance and quality of the feedback. I have repeatedly explained that we don't simply ask users what they think— we study how they use the system'. He continues to explain about feeding back results on usability issues:

It seems like a very sensitive process (...) It might have to do with the fact that the developer himself has a professional background or that he has many years of experience on his own, but it seems to be difficult for developers when someone claims that users do not understand their system (...) As a result

the developer might end up annoyed or insulted. [2U]

Further, four of six usability experts specifically express that they find some developers difficult to work with, using words like 'artists' and 'prima donnas' to suggest that some developers are unwilling to accept critique of their work.

The Developer-Project Manager Relationship

From the developers' point of view

Four developers mention how they on occasion experience that project managers do not understand or respect that creating solid code and keeping it up to date are important to developers. One developer explains how he feels pressured to cut corners to quickly solve usability problems. He explains how cutting corners will solve the problem at hand, but also dramatically weakens the code over time:

There is time pressure, right? So you cut corners, take short cuts, and do things you are not proud of professionally. But you have to in order to meet the deadline. And as a result a usability problem is reported and falls back on you (...) but you do not want to take the blame because you would like to spend a week fixing it, but you cannot. [12D]

Another developer explains a similar situation like this:

They want me to add auto layout to the forms we produce, and I explain "listen, I do not have the XML-code, so I cannot add auto layout" (...) and if I do not convince others about this, a manager, who does not get it, insists that it is done. And that is how really bad software is made. [4D]

Four interviews with developers [4D, 6D, 12D, 14D] show how they prefer usability work to be introduced earlier in the development process to avoid major changes later on. A developer explains:

When you make a new feature it has some technical aspects and some usability aspects. The problem is that you take care of all the technical aspects first, while it would be much better to do the two things in parallel. But then usability would play another part—because typically it has the critical role of providing "this is good enough, and this could be better"-comments, but if you include usability in the development process usability will have the role of "Okay, what to do about this?" [4D]

From the project managers' point of view

Three interviews with project managers describe how they sometimes struggle with convincing developers that participating in work with users will yield important information about the system. 'They do not exactly jump from joy, when they have to participate in a workshop with users', one project manager explains about some of the developers she works with, 'I do not think it is lack of will, but rather that some of them are shy and prefer to sit behind a screen' [3P]. Another project manager suggests reluctance to change the design as a reason why some developers avoid or dismiss usability work:

The developers are really skilled and experienced people (...) and have used many years on building a system to make things work. And then this young UI-designer comes along, and draws up something that do not fit anywhere. And that is really annoying and frustrating for the developers. They are rarely willing to change things. [13P]

The Project Manager-Usability Expert Relationship**From the project managers' point of view**

Three interviews with project managers [7P, 9P, 13P] suggest that usability evaluation is difficult to integrate in systems development. A major reason is that it is impossible to anticipate the outcome of tests and revise the project plan accordingly. A project manager compares usability evaluation with a bag of unknown fireworks, since it is impossible to predict what will happen once it goes off. He elaborates: 'From my point of view it can be annoying to have to include usability studies because my goal is—as quickly as possible—to reach a decision about what we need to produce' [13P]. Another project manager explains his view of the uncertainties of usability results:

There will always be the risk that the results pull the rug from under the project. Project managers fear usability tests because they might conclude that the system needs to be changed. On the other hand, they may also conclude that the solution is great—a thing we might have suspected but could not know before the test. [9P]

From the usability experts' point of view

The relationship between usability and project management differs between companies who use consultancies and those who use in-house usability experts. Consequently, the challenges also differ. Our data show that all usability experts from the consultancy companies find it frustrating to follow up on usability feedback because their job is often considered done when usability

results have been reported, or because a usability expert from outside a customer's company have little possibility to actually push decisions through [5U, 8U, 11U]. The usability experts who work in-house report how factors that influence usability, such as timing, decision-making, and planning, could be improved. To exemplify, one usability expert calls for more clarity about who can make decisions for which parts of the system [2U]. Two usability experts report that they find it difficult to include colleagues such as developers in their work, because they do not have the decision-power to book the developers' time in order to, for example, present and discuss usability findings [2U, 1U]. Finally, one usability expert explains how it—despite the project manager's good intentions—is difficult to get to do usability work early in the process [1U]. Another usability expert experiences how expenses for usability are often cut away so as to lower the price presented to the customers [4U]. These last findings suggest that usability experts feel that usability work is not prioritized as they would like.

We have elaborated on the challenges described in Table 2, and related them to relations between job roles. The findings suggest that the four challenges are important aspects when describing the work relationships between developers, usability experts and project managers. Next, we present best practices that relate to these themes.

Best Practices

In the following, we present best practices that seek to address the challenges of poor timing, usability results' lack of relevance, respect for others' job roles, and difficulty sharing important information, cf. Table 3.

Timing of usability efforts

An interview with two project managers showed how they, due to scarce resources, focus all usability attention on interdisciplinary workshops in the beginning of a project. They explain how their company has recently changed from evaluating usability later in the project to involving stakeholders, such as developers, users, customers and usability staff, at the beginning of a design process:

During the last year I have been able to see a difference in our products. Not that usability was without results before, but it was in other areas and it was not as visible (...) I am simply so happy and content about how the developers have adopted this way of thinking. It is awesome. [3P].

Best practices	Challenges addressed	N
Make early sketches and prototypes collaboratively	Timing, respect, relevance, communication	2
Share information through meetings and workshops	Respect, relevance, communication	4
Cooperatively agreeing on usability or system goals	Respect, communication	4
Use developers as informants to usability work	Respect, relevance, communication	6
Usability task force	Respect	1
Use new feedback formats such as scenarios	Communication	1
Make feedback as learning experience	Relevance, communication	1

Table 3: Best practices, and the specific challenges they address. The N-column refers to the number of interviews mentioning a specific best practice.

Because participants in the early sketching process inform the usability work with for example domain knowledge, and learn about how usability studies are done, the main benefits of moving usability work to the very beginning of a project seem not only related to timing, but also to respect and relevance.

Two developers, who also work with usability [10D], explain how they, besides initially conducting a workshop to collect and share information, invite customers to meetings during the development process. Here, they discuss and solve design issues on the spot. They describe how they sometimes hold ideas about how to solve a problem before the meeting starts, and sometimes not, but how they try to come up with a solution together with the client, and implement the solution in the prototype real time:

We treated some serious production errors during a meeting once. Even the managing director was present, and I was the technician who during the meeting made changes and updated the system. That procedure leaves a very strong impression and it takes away the argument that “this is going to be very costly” - there is always one who will argue “don’t spend any more time on that because it will get too expensive”. But if you are practically doing it real time the costs are limited. [10D]

They explain that one of the keys to their success is to insist on the participation from people with both domain knowledge and decision power. Another key is real time prototyping:

And the fact that we can show changes real time and test different solutions - that is the key. That way you can convince even the most stubborn non-believer. But you need to be prepared so that you can make changes that are immediately visible. Of course there are systems where it cannot be done, but in most cases it can. I have to admit - it was not all changes I made entirely correct, I did some dirty hacks but made it look real. But I knew that it would not take me long to make it work back home, maybe a couple of hours. [10D]

The relevance of feedback from tests

The relevance of feedback touch on issues such as the relevance of findings and recommendations, the persuasiveness or credibility of the descriptions, and how the feeding back of results is timed according to the development process. Five usability experts report how they prioritize findings to make feedback more useful. Four of these carry out the prioritization together with developers. One developer confirms the helpfulness of such a prioritized list by explaining how he and his colleagues only use the top-10 list they receive, and simply leave the more thorough report on the shelf, untouched [12D].

Another usability expert explains how he prompts developers for what they would consider appropriate findings at a given stage of development:

I have told them [developers] for example that I will not recommend any new features unless it turns out that the system does not work without them. So, in order not to scare them away I only report things that I know can be corrected. [2U]

To make the feedback more interesting one usability expert explains how he experiments with formats other than the traditional written problem description, and successfully uses scenarios, personas and illustrations as a way to make re-

sults from usability evaluations come more alive: 'It is about presenting [the results from usability evaluations] in a way that makes them an active part of the project instead of some boring report that just lies there on the shelf and collects dust' [11U].

Two project managers [3P] view feedback from a learning perspective, and explain how they successfully make developers experience problematic usability issues by not only letting them observe users, but also analyse and discuss usability matters with them:

Developers are instructed to engage in conversations with users, conduct interviews, and develop low tech prototypes. Some developers experience difficulties talking to users, and receive help and guidance from usability experts (...) This practice of self-experience has proven more effective than simply presenting and discussing usability issues at ordinary meetings. Further, involving developers in the work with users has the side effect that developers get used to thinking in terms of usability continuously and not just when the project plan dictates so. [3P].

Respect and priority

One project manager [13P] reports how his company has a usability task force based at the main office. This task force travels between local offices. To secure a high general level of usability within all products, the team has decision power over all usability issues in all projects. The project manager explains how the task force reflects positively on the smaller local usability teams because local usability teams see the existence of a high priority task force as a boost for the profession. The existence of the task force also helps raise the professional standards, and local usability experts regard the team a professional backing.

Communication and sharing of information

On the subject of sharing information, three interviews with project managers describe how workshops – understood as meetings where stakeholders collaborate to solve certain tasks – are used as a way to facilitate collaboration between usability experts and developers. Project managers explain how such workshops keep stakeholders up to date with the state of the project, and engage colleagues in other aspects of the work than solely their own. For example:

I think workshops provide developers with a better initial understanding of what it is all about. Because they have not necessarily

been a part of making the specifications (...) and if they do not know what the system is all about then I think it is really valuable for them to participate in a workshop. [3P]

Another project manager points out that working closely together also boosts team spirit and makes compromising easier: 'I think [collaboration] matters to how willing you are to change and redesign things' [9P].

In two interviews project managers explain how they use project meetings to create common references to usability, and to adjust expectations to the project. One explains how participants at project meetings each create a prioritized list of system goals. Afterwards, the individual lists are cooperatively consolidated into one, which serves as a reference for the rest of the project, helping to end discussions and make decisions:

Initially we had workshops and discussions of what is important. Is it quality? Is it usability? Is it performance? Is it response time? Is it something else? We all prioritized what we found important and we all agreed that usability was pretty high up the list. Everybody attached numbers to these topics to show what they wish to prioritize and what they want to guide the development. [9P]

This practice of collaboratively prioritising problems helps share information, and gives participants the possibility to understand their colleagues' point of view. Collaboration on prioritization is also described by a usability expert [5U] and a developer [14D]. The latter reports that being able to refer to for example usability as being an official and collaboratively agreed upon top-priority have proven very helpful when discussing and negotiating budgets with the top management.

Addressing the themes of both respect and understanding for others' work domains, a project manager describes how he brings the disciplines on a project together, and commits everyone to for instance features, prototypes, designs etc. 'People need to give something back to the project' [7P], he explains, suggesting that when people give something, for instance ideas, to a project, they experience commitment and responsibility to the project and are better motivated for working together with the other stakeholders, making compromises and otherwise contributing to the solution of problems. This experience is shared by two other project managers [3P]. However, while getting stakeholders together to overcome the challenge of different job roles is described as helpful, one project manager has a few reservations. He warns that while putting for instance usability ex-

perts and developers together in meetings make conflicting interests become clear, such experiences might also end up creating an unproductive or negative work atmosphere [7P].

Discussion

The goal of this study was to investigate what kinds of challenges developers, usability experts, and project managers experience when they collaborate on improving the usability of computer systems. We also aimed to understand which best practices are used to address such challenges, thereby attempting to develop new ideas on how to improve the collaboration between key stakeholders in systems development. Our study confirms that many of the challenges for usability work stem from tension in the relationship between job roles, as argued by for example (Gulliksen, Boivie, & Göransson, 2006). In contrast to previous work, our study investigates usability challenges specifically from the perspective of three job roles, namely developers, usability experts and project managers. The special focus on the role of the project manager and the interaction between the three job roles are perspectives rarely investigated in the present literature.

Challenges in Usability Work

Concerning the first research question, our study shows that timing, relevance, respect, and communication were all major issues for the three groups of stakeholders. These findings elaborate on results from earlier studies, such as (Rosenbaum, Rohn, & Humburg, 2000; Gulliksen, Boivie, Persson, & Hektor, 2004), by relating findings to relations amongst stakeholders. Our study suggests that these core challenges are symmetrical, in that most of them can be applied between any two job roles, like an arrow pointing back and forth. For instance, all three job roles experience challenges related to poor timing of usability work, such as feeling pressured to compromise one's professional standards. This challenge is tightly connected to project managers' experience of usability as an initiative that can pull the rug from under the project plan, and their resulting hesitation to introduce such an initiative to the project plan.

The lack of relevance of usability results relates to developers' reluctance to incorporate last minute results. Also, it seems closely related to the challenge of timing. However, lack of relevant feedback also suggests that the relevance of findings and recommendations is sometimes flawed by usability experts' lack of domain knowledge.

The challenge of respect is perhaps most clear in the developer-usability expert relationship. Both parties experience that they do not get the professional respect they deserve from colleagues. For example, developers experience usability experts' disrespect when receiving irrelevant or poorly timed usability results. Usability experts, on the other hand, interpret developers who dismiss important results as disrespecting the usability profession. Developers also feel disrespected when pressured by project managers to compromise their professional standards. While other work has pointed to usability experts struggling to get respect from colleagues (Gulliksen, Boivie, & Göransson, 2006), the observation that other stakeholders also feel ill-respected is new.

Most challenges described in this paper are related to communication. For example, learning about other professionals' job roles and goals is closely related to the challenges of respect. Sharing information about the domain seems closely related to the relevance of usability work and results. The challenge of timing relates to communication because project managers seem not to understand how usability can contribute at different stages of the project, or what to anticipate from such usability initiatives.

Let us briefly reflect on implications of our study for researching usability work. Across the literature usability work is mainly understood from the usability professionals' perspective. Accordingly, most studies report difficulties solely related to the role of the usability expert, for example (Gulliksen, Boivie, Persson, & Hektor, 2004; Gulliksen, Boivie, Persson, & Hektor, 2004). To extend this perspective, we suggest thinking in multiple perspectives, including those of developers, project managers, and top management. Exploring such perspectives may strengthen usability research. For example, several authors have argued to increase attention to developers' needs and wishes, for example (Redish, Bias, Bailey, Molich, Dumas, & Spool, 2002), and some studies have build on this argument to study the use of usability evaluation results among developers (Hornbæk & Frøkjær, 2005; Hvannberg, Law, & Larusdottir, 2007). In the present study we have discussed a new perspective, the project manager, and explored the specific related difficulties. When emphasizing multiple perspectives, we further seek to lessen the chance that a strong focus on usability experts causes us to ignore other stakeholders.

Another framework for continuing this work is seeing usability work as a cross-professional collaborative learning process. Especially our understanding of respect and communication may benefit from understanding usability in a

cross-professional context. Other studies have shown the benefits of working closely together in cross-professional settings when it comes to learning about other job roles and other professionals' point of view (Bødker & Buur, 2002; Furniss, Blandford, & Curzon, 2007). In this frame understanding professionals as human beings with individual values, strengths and weaknesses might also help us explore why collaboration on usability issues is complex and difficult. The view that stakeholders are also individuals who work within social relationships with customers and colleagues is not new, see for example (Furniss, Blandford, & Curzon, 2007; Iivari, 2006). However, stories that tell us that 'loud' individuals have a better success rate in some companies, or how personal and professional respect seem to rely on social skills (Iivari, 2006) suggest that we do not give the role of the individual enough attention. We do believe that job roles are of importance when it comes to collaborating to improve usability, but when it comes to collaboration we might also need to look at how different individuals support each other. Or do not. In this respect Furniss et al. (2007) have already identified negotiation skills as having huge importance when it comes to collaborating efficiently, and we suggest looking into related social traits such as empathy, humour and diplomatic skills.

Best Practices

Concerning the second research question, the study shows how best practices already address some or more of the challenges. For example, moving all usability initiatives to the beginning of a project is a way of dealing with the challenge of timing. To prioritize project goals collectively is a way to share information about professional goals, and addresses the need for better communication. Using developers as informants is a way to show and build respect, in addition to improving the relevance of the results. Such an approach might also help improve developers willingness to carry out recommended fixes, as psychological studies have shown (Benton, Kelley, & Liebling, 1972; Schindler, 1998).

Looking at the challenges and the best practices uncovered in this study, our advice to usability practitioners is as follows:

Do not present usability findings in the last minute to developers. Find ways to do the work earlier such as using rapid prototyping or early workshops or postpone initiatives to the beginning of a second round of development.

- Give relevant feedback. Engage colleagues in the usability work to ensure that findings and recommendations rest on solid knowledge

about what can be fixed, how, and when.

- Show respect for other professions. Do not dismiss colleagues and their viewpoints simply because they differ from your own professional goals and work practices. Understand that your goals might conflict with colleagues' professional goals.
- Share knowledge. Engage colleagues who have a stake in your work, share viewpoints, discuss, and join efforts to set and prioritize tasks and goals.

In the present study best practices are mostly tuned towards learning, such as learning about other stakeholders' professional standards, and collaborating, such as jointly agreeing on system goals, such as described by (Mayhew, 1999). To get a better understanding of how usability work can be understood as a collaborative learning process, we suggest looking deeper into how such processes are supported or impeded in the current work practice.

Dumas (in Redish, Bias, Bailey, Molich, Dumas, & Spool, 2002) has argued that the personal relationship between developer and usability expert might be the most important factor for usability's success, more important than, for example, how usability results are fed back to developers. Others have made similar observations on the importance of human relationships, such as the relationship between usability expert and customer, users and stakeholders, and so on (Bennet & Karat, 1994; Furniss, Blandford, & Curzon, 2007; Wixon & Wilson, 1997). In fact, Bennet and Karat (1994) argued that finding ways to facilitate collaboration between stakeholders to usability is a most urgent matter for HCI research. Because learning and collaboration seems to be such a key concept when designing usable systems, we suggest investigating the perspective of usability as a human activity rather than as a matter of methods and procedures. The Participatory Design tradition (Bødker & Buur, 2002) reflects this perspective but focuses mostly on the beginning of the development process, where much has the form of sketches. Understanding usability work in the perspective of human activities, rather than processes and methods, will perhaps help researchers and practitioners bring the focus on for example collaboration and learning beyond the sketching phase of development, and into other parts of development relevant for usability, namely evaluation.

Next, we briefly review four papers that in various ways deal with how usability practitioners work together with other stakeholders in the industry. To better understand how our study contributes

to the general understanding of the cooperative aspects of usability work, we then relate these papers to the present study.

Discussion of four related papers

Furniss et al. (2007) aim to describe what happens in industrial practice between stakeholders and usability professionals. Their work shows that customers have much influence on usability work, and that this influence increases when there is tension between the customer and the usability expert. They see usability work as a collaborative effort and show how personal relations are important for the customer-usability practitioner relationship. Because usability work is no one-man show, they call for a better understanding of how individuals and professionals can cooperate to produce valuable usability work.

Gulliksen et al. (2006) have studied usability professionals on an individual level to investigate which success factors and obstacles they encounter. They conclude that individual background and experience can improve or impede the quality and success of usability work as well as organisational characteristics and stakeholders' attitudes towards usability. The paper is written from the perspective of the usability practitioner and mostly deals with this role: what practitioners do, how they do it, and the quality and results of their work. Since the paper is based on studies of and interviews with usability practitioners, the description of this job role and its challenges seems perhaps one-sided. For example, we learn that a great portion of usability practitioners consider themselves well-informed about the system domains they work with, while our study suggests that developers may disagree. Other issues such as respect or the importance of being on good terms with the project manager, is also discussed in the paper. The paper lists problems and challenges for usability practitioners' work, but does not proceed far into why such challenges exist and hence only superficially into how to address them. For example, the paper argues that insufficient authority is a problem for usability practitioners, but only briefly explores why that might be (except that it is an 'attitude problem' in systems development at large).

While Gulliksen et al. (2006) have organisation as one of many topics, Iivari (2006) presents a case study entirely on the relationship between organisational and usability work cultures. Iivari's study mostly concerns organisational matters such as responsibility and power structures in different organisational cultures. However, it touches on issues related to the present study. For example, the paper mentions conflicts between colleagues on a project and argues that they may be caused

by strong personalities and an organisational culture where loud individuals succeed. Iivari's paper also points to other issues similar to the ones discussed in this paper: how project management is often considered insufficient, how some stakeholders are considered very sensitive about their work, how lack of respect can be a problem between colleagues, how the timing of usability initiatives are often bad, and how it may seem difficult to include usability work in project plans.

Bødker and Buur (2002) discuss how to facilitate better knowledge sharing and collaboration on design, and describe a number of best practices. The main topic of their paper is how to improve design through better collaboration in a setting called the Design Collaboratorium. They present a point of view different from our study, which aims at investigating which collaborative challenges different job roles experience, and how one may improve collaboration by addressing these challenges in different ways. The work with the Design Collaboratorium seems based on earlier research findings that showed how 'usability issues were brought into the design process too late and with too little to say' (Bødker & Buur, 2002). The paper by Bødker and Buur does not identify any reasons for why usability enters the design process too late, or what the more specific consequences are – besides it having 'too little to say'. Also, the Design Collaboratorium seems best applied relatively early in the design process, and is perhaps best suited for certain types of systems. It also demands quite a lot of planning and may thus run into the exact same problems with project managers that usability work does, namely that they do not know when or how to integrate the exercise into the project plan.

If we compare the four papers with the study we have conducted, our study seems to add to several of the key findings in the papers above. Furniss et al. (2007) look at relationships between usability practitioners and a group defined only as 'customers'. Some of the stakeholders in our study consider themselves 'customers', but are also very aware of their profession and job role. While Furniss et al. (2007) argue for the importance of understanding groups of customers or usability practitioners as also being individuals with individual skills; we argue that those groups should also be understood as consisting of people with different job roles. Adding the perspective of job roles to the one of individuals is important because our study shows that individuals who hold the same job role share challenges. However, based on our experiences from the present study we are convinced that the focus on individual skills and characteristics such as empathy, humour or diplomacy is also of great importance

to cross-professional collaboration and should be studied further.

While Gulliksen et al. (2006) describe usability work and relations from the view of the usability practitioner, and Iivari focus on organisational culture, our study aims to investigate and understand three job roles, and not particularly take the stand of the usability practitioners.

The four papers all point to problems that are related to the challenges identified in this paper. Still, we provide some new explanations of why such problems and challenges occur. For example, Furniss et al. (2007) argue that usability work include making difficult pragmatic decisions regarding for example budgets and deadlines. Our focus on job roles suggests that these difficult choices mainly lies with the project managers, and not so much the usability practitioners, as one might expect. Also, when Furniss et al. discuss the matter of tension between customers and usability practitioners, and Iivari (2006) points to conflicts between different colleagues on a project, we can provide examples on how this is manifested in the daily work between job roles. We argue that tension in relationships is mostly related to the relationship between developers and usability practitioners. The focus on roles also suggests why tension may occur, since many participants in our study refer to a lack of respect between these two roles. To give due credit, Iivari offers interesting points on the question of what builds personal and professional respect in different types of companies, for example how excellent social skills help build respect amongst co-workers.

Generally, the papers only deal with concrete best practices in a limited fashion. The exception is Bødker and Buur (2002). Nevertheless, they run the risk of presenting work procedures that are too ambitious or complicated to be easily used in the industry. The best practices presented in our paper may seem less ambitious than those of Bødker and Buur, but they are also less risky viewed from a project manager's point of view. Accordingly, they may stand a better chance of being used.

While all papers discuss challenges for usability work from different perspectives such as customers or organisational culture, they only sporadically investigate why such challenges exist. Our study suggests that the challenges people encounter when working together to improve usability can be understood from the perspective of job roles, and that usability work for these reasons is best explained as a collaborative cross-professional learning process.

Limitations of results

Since this study is conducted as interviews the findings may be the result of a subsequent rationalization on behalf of some of the interviewees. Consequently, this study investigates the participants' perceived challenges. In-situ observations of the interactions between stakeholders might provide us with a better understanding of whether perceived challenges differ from actual challenges and identify unsaid practices and barriers.

Investigating the perceived challenges in the relationship between three groups of stakeholders only addresses parts of a very complex problem. We would like to investigate if stakeholders who have more than one job role, such as a project manager with a background in usability studies, have different perspectives than stakeholders with only one job role. Also, getting hold of a broader sample of informants might provide new results. For example, one may speculate whether members of special interest groups (SIGs) are different from professionals that are not SIG members, or whether the developers that were introduced by SIG members were perhaps more experienced with and interested in usability work than developers in general.

Similarly, we might expect that investigating other key stakeholders, for example the top management, could be relevant to understanding especially the challenge of timing, but perhaps also to the challenge of respecting colleagues' professional goals.

Iivari (2006) and Gulliksen et al. (2006) argued that usability work is also influenced by various organisational characteristics. This may very well also be the case for the relationships amongst colleagues. However, gathering thorough organisational characteristics has not been a focus of this study. As a result, challenges that relate to cross-professional relationships in various organisational settings need to be understood before we can draw any generalizable conclusions about the complex pattern of challenges for usability work.

Since this study primarily focuses on evaluation work we might have limited ourselves by only allowing stakeholders to discuss best practices in relation to evaluation. As a consequence, we might be guilty of ignoring other best practices such as those related to the design of an underlying architecture that can easily be changed. In sum, further work should aim to describe challenges in a broader perspective taking into account that usability work takes place in a complex organisational setting between several groups of stakeholders and that evaluation work is only a part of a series of tasks that influence usability.

Conclusion

Many seem to consider usability work a well-integrated and well-understood part of software development. However, it still does not seem to impact the development of software as much as usability professionals desire. Our study of the relationships between developers, usability experts, and project managers suggests that looking into the interaction between these stakeholders can help us better understand why.

The study shows that challenges related to the timing of usability work, the lack of relevance of usability results, disrespect for others' job roles and goals, and difficulties in sharing and getting important information are key challenges for the cooperation between the participants. These challenges have been known for many years to impede usability work. The surprising finding is that despite the implementation of clever best practices and work-arounds those well-known challenges are still reported to be the top show-stoppers for effective usability work. We report best practices such as joint sketching or collaboratively deciding on project goals as ways to address these challenges. We propose four overall guidelines to facilitate better relationship and interaction between developers, usability experts and project managers, and suggest looking further into how such guidelines can be used in different work contexts.

Also, we recognise that difference in job roles cannot explain every single problem with cross-professional collaboration. We need to acknowledge that personal relationships between individuals also have a major impact on how well people work together. In this respect Furniss et al. mentions negotiation skills, and we suggest empathy, humour and diplomatic skills as being worth studying in the future.

Gulliksen et al (2006) conclude their paper by summing up a 'frivolous' description of what a usability practitioner needs to succeed:

You need systems developers that are brilliant programmers and ready to put in as much time as required to do as you bid, and at the same time willing to make numerous modifications to their solutions in order to accommodate the changing requirements inherent to systems development, without complaint. You need a client that is committed to user-centred design, willing to spend unspecified amounts of money on your development project. And you need users that are willing and able to spend unspecified numbers of hours with the project in various

analysis, design and evaluation activities. As well as being at your beck and call, at any time of the day to answer all the detail questions that are inevitable throughout the entire course of the project. (Gulliksen, Boivie, & Göransson, 2006)

Perhaps we may offer an equal frivolous summary of how cross-professional collaborations on usability work succeed:

You need systems developers that are always happy to receive usability critique, will gladly change the code at any point in time, and passionately engage in usability work. You need usability experts with detailed knowledge of the system domain, the system's code, and the progress of the development, who only suggest top-relevant design changes, and do so with perfect timing. You need project managers who satisfactorily involve everybody in the planning of the project, give top-priority to usability at all times, while meticulously following up on all recommendations, and still leave room for developers to follow their own professional standards. And you need all these people to hold the utmost respect for each other professionally and personally, possess excellent communication and diplomatic skills, and be thrilled with joy about working together at all times.

While this description may not be a serious attempt to outline how successful cooperations are built, it does capture the challenging nature of getting cross-professional collaborations to succeed. And in this study we have only looked at three job roles, while usability experts' relationship with for example top management and marketing is still to be studied.

References

- American National Standards Institute. (2001). The Common Industry Format (ANSI/NCTS-354-2001) .
- Bennet, J., & Karat, J. (1994). Facilitating effective HCI design meetings. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Celebrating Interdependence, 198-204.
- Benton, A. A., Kelley, H. H., & Liebling, B. (1972). Effects of Extremity of Offers and Concession Rate on the Outcomes of Bargaining. Journal of Personality and Social Psychology, 24, 73-83.
- Bødker, S., & Buur, J. (2002). The design collaboratorium - a place for usability design. ACM Transactions on Computer-Human Interaction (TOCHI) .
- Bødker, S., & Krogh, P. M. (2001). The interactive design collaboratorium. Proceedings of the Interact 2001 .

- Boland, R. (1978). The process and product of system design. *Management Science*, 24, 887-898.
- Chattratchart, J., & Brodie, J. (2004). Applying user testing data to UEM performance metrics. CHI '04 Extended Abstracts on Human Factors in Computing Systems, Vienna, Austria, 1119-1122.
- Chi, M. (1997). Quantifying Qualitative Analyses of Verbal Data: A Practical Guide. *The Journal of the Learning Sciences*, 6, 3, 271-315.
- Coble, J., Karat, J., & Kahn, M. (1997). Maintaining a focus on user requirements throughout the development of clinical workstation software. *Proceedings of the ACM Conference on Human Factors in Computing*, 170-177.
- Cockton, G., Woolrych, A., & Hindmarch, M. (2004). Reconditioned merchandise: Extended structured report formats in usability inspection. CHI '04 Extended Abstracts on Human Factors in Computing Systems, Vienna, Austria, 1433-1436.
- Dumas, J. (1989). Stimulating change through usability testing. *SIGCHI Bulletin*, July, 1989, 21, 1, 37 - 44.
- Dumas, J., & Redish, J. (1999). A practical guide to usability testing. Oregon, USA, Intellect Books.
- Ehn, P. (1992). Scandinavian design: On participation and skill. In P. Adler, & T. Winograd, *Usability: Turning Technology into Tools*, 96-132. New York, Oxford University Press.
- Ehn, P., & Sjögren, D. (1991). From system descriptions to scripts for action. In J. Greenbaum, & M. Kyng, *Design at Work - Cooperative Design of Computer Systems*, 241-268. Hillsdale, New Jersey, Lawrence Erlbaum.
- Furniss, D., Blandford, A., & Curzon, P. (2007). Usability Work in Professional Website Design: Insights From Practitioners' Perspectives. In E. Law, E. Hvannberg, & G. Cockton, *Maturing Usability: Quality in Software, Interaction and Value*, 144-167, Springer London.
- Greenbaum, J., & Kyng, M. (1991). *Design at Work: Cooperative Design of Computer Systems*. Hillsdale, New Jersey, Lawrence Erlbaum Associates.
- Gulliksen, J., Boivie, I., & Göransson, B. (2006). Usability Professionals - Current Practices and Future Development. *Interacting with Computers*, 18, 568-600.
- Gulliksen, J., Boivie, I., Persson, J., & Hektor, A. L. (2004). Making a Difference - a Survey of the Usability Profession in Sweden. *Proceedings of Nordichi 2004*, 207-215.
- Hertzum, M., & Jacobsen, N. (2001). The Evaluator Effect: A Chilling Fact About Usability Evaluation Methods. *International Journal of Human-Computer Interaction*, 13, 421-443.
- Hornbæk, K., & Frøkjær, E. (2005). Comparing usability problems and redesign proposals as input to practical systems development. *ACM Conference on Human Factors in Computing Systems*, 391-400.
- Hvannberg, E. T., Law, E., & Larusdottir, M. K. (2007). Heuristic evaluation: Comparing ways of finding and reporting usability problems. *Interacting With Computers*, 19, 2, 225-240.
- Iivari, N. (2006). 'Representing the User' in Software Development - a Cultural Analysis of Usability Work in the Product Development Context. *Interacting with Computers*, 18, 635-664.
- Ives, B., & Olson, M. (1984). User involvement and MIS success: A review of research. *Management Science*, 30, 586-603.
- Johnson, D., & Johnson, F. (1990). *Joining Together*. Englewood Cliffs, NJ, Prentice Hall.
- Karat, C., Campbell, R., & Fiegel, T. (1992). Comparison of Empirical Testing and Walkthrough Methods in Usability Interface Evaluation. *Proceedings of CHI'92*, 397-404.
- Kennedy, S. (1989). Using video in the BNR usability lab. *SIGCHI Bulletin*, 21, 2, 92-95.
- King, W. R., & Rodriguez, J. J. (1981). Participative design of strategic decision support systems: An empirical assessment. *Management Science*, 27, 717-726.
- La Fasto, F., & Larson, C. (2002). *When Teams Work Best*. Thousand Oaks, CA, Sage Publications.
- Law, E., & Hvannberg, E. T. (2004). Analysis of strategies for improving and estimating the effectiveness of heuristic evaluation. *Proceedings of the Third Nordic Conference on Human-Computer Interaction NordiCHI '04*, 241-250.
- Madsen, K. H., & Petersen, M. G. (1999). Supporting collaboration in multi-media design. *Human-Computer Interaction - INTERACT'99*, 185-190.
- Mayhew, D. (1999). *The Usability Engineering Lifecycle. A Practitioner's Handbook for User Interface Design*. San Francisco, CA, Morgan Kaufmann.
- Mills, C. (1987). Usability testing in the real world. *SIGCHI Bulletin*, 18, 67-70.
- Nayak, N., Mrazek, D., & Smith, D. (1995). Analyzing and communicating usability data. *SIGCHI Bulletin*, 27, 1, 22-30.
- Redish, J., Bias, R., Bailey, R., Molich, R., Dumas, R., & Spool, J. (2002). *Usability in practice: Form-*

- tive usability evaluations - Evolution and revolution. ACM Conference on Human Factors in Computing System, Minneapolis, Minnesota, 885-890.
- Robey, D., & Farrow, D. L. (1982). User involvement in information system development. A conflict model and empirical test. *Management Science*, 28, 73-85.
- Rosenbaum, S., Rohn, J. A., & Humburg, J. (2000). A toolkit for strategic usability: Results from workshops, panels and surveys. *Proceedings of the ACM CHI 2000 Conference on Human Factors in Computing Systems*, 1, 337-344.
- Rubin, J. (1994). *Handbook of Usability Testing: How to Plan, Design and Conduct Effective Tests*. New York: John Wiley & Sons inc.
- Schell, D. (1986). Usability testing of screen design: Beyond standards, principles, and guidelines. *Proceedings of the Human Factors Society 30th Meeting*, Santa Monica, CA, 1212-1215.
- Schindler, R. M. (1998). Consequences of Perceiving Oneself As Responsible for Obtaining a Discount. *Journal of Consumer Psychology*, 7, 371-392.
- Strauss, A., & Corbin, J. (1998). *Basics of Qualitative Research - Techniques and Procedures for Developing Grounded Theory*. California, Sage Publications.
- Uldall-Espersen, T., & Frøkjær, E. (2007). Usability and software development: Roles of the stakeholders. *Proceedings of HCI2007*, July 22-27, Beijing, China, 642-651.
- Uldall-Espersen, T., Frøkjær, E., & Hornbæk, K. (2007). Tracing Impact in a Usability Improvement Process. *Interacting With Computers*, 48-63.
- Winer, M., & Ray, K. (1994). *Collaboration Handbook: Creating, Sustaining, and Enjoying the Journey*. Lafond, St. Paul, MN, Amherst H. Wilder Foundation.
- Wixon, D., & Wilson, C. (1997). The usability engineering framework for product design and evaluation. In M. Helander, T. Landauer, & P. P., *Handbook of Human Computer Interaction*, 653-688, North-Holland, Elsevier Science.

Exploring the Value of Usability Feedback Formats⁵

Mie Nørgaard

University of Copenhagen
Universitetsparken 1
DK- 2100 Copenhagen
mien@diku.dk

Kasper Hornbæk

University of Copenhagen
Universitetsparken 1
DK- 2100 Copenhagen
kash@diku.dk

Abstract

The format used to present feedback from usability evaluations to developers affects whether problems are understood, accepted, and fixed. Yet, little research has investigated which formats are the most effective. We describe an explorative study where three developers assess 40 usability findings presented using five feedback formats. Our usability findings comprise 35 problems and 5 positive comments. Data suggest that feedback serves multiple purposes. Initially, feedback must convince developers about the relevance of a problem and convey an understanding of this. Feedback must next be easy to use and finally serve as a reminder of the problem. Prior to working with the feedback, developers rated redesign proposals, multimedia reports, and annotated screen dumps as more valuable than lists of problems, all of which were rated as more valuable than scenarios. After having spent some time working with the feedback to address the usability problems, there were no significant differences among the developers' ratings of the value of the different formats. This suggests that all of the formats may serve equally well as reminders in later stages of working with usability problems, but that redesign proposals, multimedia reports, and annotated screen dumps best address the ini-

⁵This paper is in press for The International Journal of Human-Computer Interaction, 2008.

tial feedback goals convincing developers that a usability problem exists and of conveying an understanding of the problem.

Introduction

Since usability studies became established as an important activity in systems development, the effectiveness of usability evaluation methods has been investigated thoroughly, see for instance (Jeffries, Miller, Wharton, & Uyeda, 1991; Sears, 1997; John & Marks, 1997). The literature focuses on comparing usability evaluation methods, but tends not to focus much on how the evaluation results are fed back to a design team, though see Dumas, Molich, and Jeffries (2004), and Hornbæk and Frøkjær (2005). This is unfortunate since one goal of usability evaluation is to improve systems. To reach this goal, evaluations must move beyond solely listing usability problems and help developers decide which usability problems to fix and how to fix them.

The Oxford English dictionary (askoxford.com) describes feedback as: 'Information given in response to a product, performance etc., used as a basis for improvement'. According to this definition, feedback needs to fulfil certain requirements to be successful. The receiver must understand the feedback, and the feedback needs to facilitate the solving of a given problem. To do this, the feedback needs to be convincing. Consequently, an evaluator about to feed back results to a development team faces at least two challenges. First, developers may not be easily convinced about usability problems, either believing that the system is great as it is or that users eventually will learn to use it (Kennedy, 1989; Seffah & Andreevskaja, 2003). Second, developers might not be hostile to changes, but simply find it difficult to understand a usability problem because it is vaguely described (Dumas, Molich, & Jeffries, 2004). How evaluators tackle these two challenges can influence the evaluation's impact dramatically.

The present explorative study aims to describe the practical use of different feedback formats and thus identify how we more successfully can feed back usability findings to developers. The study investigates how five feedback formats are used and assessed by developers. These formats represent different ways by which an evaluator might deliver usability results to developers. The results suggest that developers initially value information in addition to the problem description, such as videohighlights, contextual screen dumps, and redesign proposals. After having worked with the feedback, the differences between feedback

formats diminish. We argue that these results are important for usability practitioners for choosing amongst feedback formats and for researchers as a help to understand how feedback is used.

Related work

Related work may be divided into two categories; one characterizing feedback practices, and another concerned with feedback research. Below we discuss the two categories in turn.

Feedback practices

The literature on feedback from usability work suggests that merely providing a description of the usability problems is insufficient, and it comprises attempts to improve feedback's persuasiveness and to facilitate the fixing of the problems. Accordingly, feedback from usability evaluations may include descriptions of a problem's severity (Dumas, 1989; Kennedy, 1989; Coble, Karat, & Kahn, 1997; Hornbæk & Frøkjær, 2005), the context of a problem (Kennedy, 1989; Nayak, Mrazek, & Smith, 1995), redesign proposals (Jeffries, 1993; Nayak, Mrazek, & Smith, 1995; Dumas, Molich, & Jeffries, 2004; Hornbæk & Frøkjær, 2005), and underlying causes of problems (Dumas, 1989). Practitioners and researchers also agree on the persuasive power of developers seeing users interact with the system (Schell, 1986; Mills, 1987; Dumas, 1989; Redish, Bias, Bailey, Molich, Dumas, & Spool, 2002).

Most of the related literature discusses feedback in terms of isolated report features, such as the use of redesign proposals. We next describe how such features might be put together to comprise different feedback formats.

An informal survey conducted in an online forum for usability practitioners suggested that a usability report containing a list of problems is perhaps the most common way to feed back usability results to developers. In relation to problem lists, researchers such as Molich (in (Dumas, Molich, & Jeffries, 2004) have argued for the importance of presenting positive comments together with the usability problems. Molich argued that developers find it valuable to know which parts of a system that work well, and that combining positive and negative criticism is the most pedagogical way to present feedback. Problem lists may describe each usability problem with a short text and a severity rating. Severity ratings may be used to generate a top 10-list of the most critical problems so as to help developers prioritize their work and reduce the number of problems reported (Dumas, 1989; Nielsen, 1993; Nayak, Mrazek, & Smith, 1995; Re-

dish, Bias, Bailey, Molich, Dumas, & Spool, 2002).

Nayak et al. (1995) described a format consisting of screen dumps annotated with recommendations for usability improvements. This feedback format aims at providing developers with example-based references to support the development process and emphasizes that the description of the usability problem is linked closely to the system context in which it occurred.

Nayak et al. (1995) also described multimedia presentations as interactive documents that mix descriptive text with video highlights, pictures, and graphics. The information is linked in a structure similar to web pages and presents information on demand. The multimedia presentation format is created with the expectation that graphical and video input increases the feedback's quality and persuasiveness.

As an elaboration of the video highlights, Dumas and Redish (1999) discussed a professional video production that resembles video productions as known from TV, including a narrator, voiceover, and examples from the test. Dumas and Redish suggested that usability feedback presented in a well-known professional format might increase the feedback's persuasiveness and might also be more enjoyable to work with than a list of problems. In contrast, professional video productions are expensive to produce and time-consuming to use. This may be the reason why no practitioners in our survey mentioned using professional video production as a feedback format.

In the literature, redesign proposals are referred to as constructive input that provides developers with ideas for fixing problems (Jeffries, 1993; Hornbæk & Frøkjær, 2005). Redesign proposals can include a brief summary of the redesign, a justification of the proposed design, an explanation of the interaction and design decisions in the redesign, and illustrations of how the redesign works (Hornbæk & Frøkjær, 2005). The use of redesign proposals is thought to inspire and help developers solve the reported problems, while the use of justifications and explanations is thought to improve the format's persuasiveness.

Scenarios are stories describing a user's goal, system interaction, and contextual factors that relate to product use (Rosson & Carroll, 2002). They build upon results from real users (Nayak, Mrazek, & Smith, 1995) or task analysis (Nielsen, 1993). Scenarios are only rarely mentioned as a way to provide feedback. This surprises, since their focus on the context of use and user behaviour might provide developers with valuable information about a usability problem. The human-centered story is one type of scenario that uses dialogue

and directly describes the characters' emotions and motivations (Strøm, 2003). Despite its focus on contextual information, human-centered stories have yet to be used for describing usability problems.

The work on textual feedback methods aside, the value of oral feedback is not to be overlooked. Oral feedback is a means to describe and initiate a dialogue about results (Kahn & Prail, 1993; Butler & Ehrlich, 1993; Dumas & Redish, 1999), and is appreciated for its power to clear up potential misunderstandings in an engaging interaction between evaluator and developer. However, oral feedback may be quickly forgotten and needs to be documented to be useful after a period of time.

To sum up, the reviewed literature on usability feedback and feedback formats focuses on how to convince a receiver about the relevance of usability problems and on how to facilitate the fixing of problems, for instance by sketching design ideas or presenting top-priority problems. However, it seems to understand written feedback as a product that plays one continuous role during the development of a system. Our study broadens this view by suggesting that the role of feedback changes over time.

Feedback research

Few studies have investigated the use of different feedback formats. Accordingly, researchers need to pay more attention to how developers use and assess various types of feedback from usability evaluation. Below we discuss studies that have done this.

Cockton (2006) recently argued that usability studies have moved from looking at evaluations as merely generators of problem lists to dealing with problems' impact. One line of work in this direction concerns downstream utility (John & Marks, 1997; Hornbæk & Frøkjær, 2005; Law, 2006), that is, the effectiveness with which a solution to a usability problem is implemented. In a study of downstream utility, Law (2006) suggested 'credibility' as a key factor for effective feedback and described how developers need to be convinced about, for example, the evaluator's expertise before acknowledging the feedback. She suggested that the persuasive power of feedback lies in providing the developers with information about the severity of the usability problem, problem frequency as well as elaborate and accurate problem descriptions. Good feedback seems also to include redesign proposals and an estimated fixing effort (Law, 2006). Other work (Nørgaard & Høegh, 2008) points to the quality of the arguments and their ability to engage as important

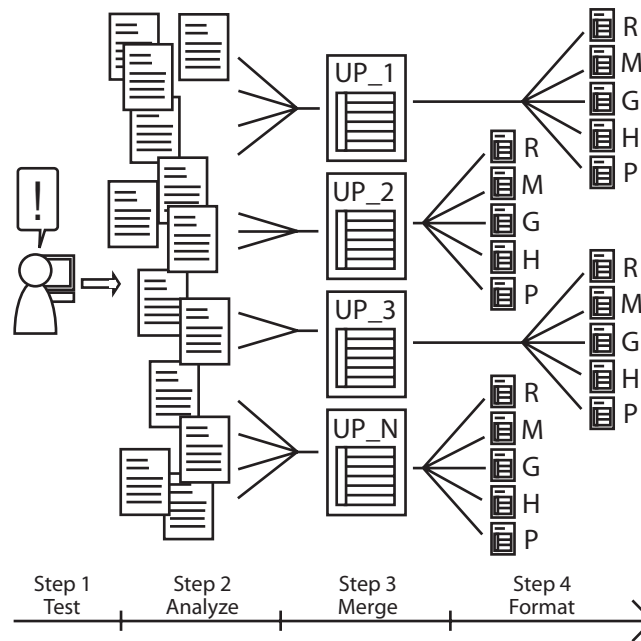


Figure 1: The figure shows the first four steps of the study, namely how the usability test (step 1) was followed by analysis of 75 usability problems (step 2) and the merging of these into 40 usability problems (step 3). Finally five feedback items were constructed for each usability problem, a total of 200 items (step 4). The figure refers to a usability problem as ‘UP’.

factors for whether the feedback is acknowledged. The views of both Law (2006) and Nørgaard and Høegh (2008) reflect an understanding that feedback, in addition to being persuasive, should facilitate developers’ work in more ways than by simply describing the problem.

A recent special issue of the *International Journal of Human-Computer Interaction* called for more research on the ‘various forms of feedback in which the results of usability evaluation is presented to developers’ to examine persuasiveness and impact (Hornbæk & Stage, 2006). This explorative study aims at investigating how well various formats convince developers and help them understand a usability problem. While most work seem to overlook that time and use are important factors for how feedback formats are valued, this study aims to broaden our understanding of what makes a good feedback format by investigating these issues. The short-term goal is to get better knowledge of how evaluators should present their feedback to developers for it to be understood and used. The long-term goal is to make evaluation a more powerful player in software development, something only rarely the case today (Hornbæk & Stage, 2006).

Method

To identify effective ways of providing feedback, we investigated how five feedback formats influenced usability work in a Danish company. This setup was chosen because it allowed us to study the work on a running system in realistic settings and provided an opportunity to investigate how developers assess feedback when first presented to them, and how they rate the same feedback once they have worked with it for some time.

The study consisted of eight steps. The system was tested, the usability problems identified, analysed, and merged into groups. Then, the problem descriptions were expressed in five feedback formats, and developers rated these on five questions. The developers then worked with the feedback, re-rated it, and were finally interviewed about their ratings. The eight steps are described in detail below (see also Figure 1 and 2).

The study was designed to investigate how different feedback formats would convince developers about the relevance of problems and provide them with an understanding of these (see questions 1-5 in Table 1). We also expected the study to provide qualitative data on how to improve feedback from evaluations to developers. We hypothesized that the ratings of the first impressions of the feedback and the ratings after it had been

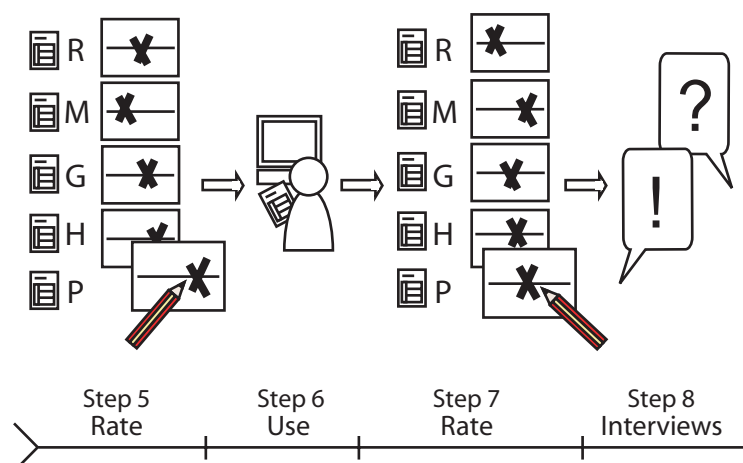


Figure 2: The figure shows steps five to eight. The developers rated the usefulness of the 200 feedback items (step 5). They then worked with 40 items (step 6) and re-rated these after completing their work (step 7). Finally, the developers were interviewed about the use of the formats (step 8).

used would vary, since working intensively with a format might bring the developers to appreciate certain qualities of a format. In the following we use the term pre-use when referring to ratings of first-hand impressions and the term post-use when referring to ratings that were given after the feedback had been used.

The company in which the study took place is Jobindex, a non-hierarchally organised company with 37 employees who provide web based services related to job searching. The three developers who participated in the study composed the development team concerned with systems development, and were accustomed to receiving feedback from usability evaluations. The developers are referred to as Dev1, Dev2 and Dev3.

Step one—testing the system

A think aloud test of the system comprised six test sessions, and followed the guidelines of Dumas and Redish, (1999). Jobindex identified the test's focus, and approved the tasks for the test. The test sessions were recorded on digital video using a webcam and Tech Smith's Morae software. The goal of the test was to sample a set of usability findings for the study, not to uncover every issue in the system.

Step two—analysing the results

To identify usability issues, two evaluators discussed and analysed the test results immediately after each test session, as recommended by Nør-

gaard and Hornbæk (2006). After the six sessions, usability findings were consolidated. Problems were described with a title, a description of the problem, a severity rating, details on the context in which the problem occurred, and one or more redesign ideas. As recommended by Dumas and Redish (1999) we included positive findings. These were described with a title, a description of the positive finding and the context in which it occurred. At the end of step two, 75 usability findings had been described: 67 usability problems and eight positive findings.

Step three—merging usability findings into 40 groups

To eliminate doublets, the 75 usability findings were merged into groups of related problems. The usability findings were merged by rough similarity until 40 groups emerged. This limit was set to ensure that the developers would get experience in working with each feedback format during step six in the study. The result of the merging was 35 usability problems and 5 positive findings. Since the positive findings are not relevant to the present paper they are ignored here, and we refer to the usability findings as usability problems during the rest of the paper.

Step four—turning the findings into feedback items

We chose to investigate five feedback formats that represent different approaches to providing feed-

		Q1: How useful is the feedback item to your work on jobindex.dk? (not useful/very useful)	Q2: How well does the feedback item help you understand the problem? (poorly/very well)	Q3: How well does the feedback item help you solve the problem? (poorly/very well)	Q4: How convinced are you that this is a problem? (poorly/very well)	Q5: How easy is the feedback item to use in your work on Jobindex.dk? (poorly/very well)
Problem List	M SD	58.1 13.0	59.1 16.4	28.5 12.7	44.6 15.3	36.0 15.1
Screen dump	M SD	61.9 11.1	67.6 11.8	45.2 15.7	50.6 14.4	44.7 14.4
Multimedia presentation	M SD	60.0 9.8	69.0 14.7	43.1 17.5	54.0 13.8	43.0 15.7
Redesign proposal	M SD	63.9 11.3	72.8 10.2	50.8 17.0	51.0 13.8	46.4 14.7
Scenario	M SD	38.8 12.6	41.3 14.3	18.4 8.8	32.7 13.5	26.8 8.9
F-test		$F(4,170)=26.68, p<.001$		$F(4,170)=28.87, p<.001$	$F(4,170)=11.73, p<.001$	
			$F(4,170)=29.57, p<.001$		$F(4,170)=12.46, p<.001$	
Tukey HSD post hoc test		SCE<PRO, MUL,SCR,RED	SCE<PRO< RED,MUL,SCR	SCE<PRO< RED,MUL,SCR	SCE<PRO< RED,MUL,SCR	SCE<PRO< RED,MUL,SCR

Table 1: Average pre-use ratings of question Q1-Q5. A format listed in two significant groups in the Tukey HSD post hoc test column (such as SCR in Q3) is neither significantly different from one group or the other. The rating spans the numbers 1-100, 100 being the best rating.

back from evaluations. As mentioned above, formats were chosen based on our literature review and on an informal survey amongst practitioners.

The study investigates the following formats: The list of problems (PRO, Figure 3) consisted of a description and a severity rating of the usability problems. Severity was rated according to a five-step scale (Dumas, 1989): Level 1 prevented users from performing or completing a task; level 2 caused significant frustration; level 3 caused some frustration; level 4 did not significantly affect usability; level 5 identified a problem that was only relevant for product enhancement in a following release. The format was included in the study since it is a common way to present usability feedback and since it can be produced at low cost. The problem list took approximately half a day to prepare.

The screen dump format (SCR, Figure 4) consisted of screen dumps annotated with information about where the usability problem occurred, a brief description of the usability problem, and a

description of one or more possible solutions. The screen dump format was included in the study because it could be produced at a fairly low cost and because it focussed primarily on presenting the context of the problem and only briefly touched upon possible redesign issues. The screen dump format took approximately one day to prepare.

The multimedia presentation (MUL, Figure 5) consisted of linked html-documents describing the problem, a video with examples of user interaction, a description of one or more solutions, a graphical illustration of severity, illustrative drawings that helped skim the content, illustrations of both problem and possible solution, and a short explanation of the illustrations. This format was included in the study because it follows the recommendations to let developers see real users interact with the system. Also, the multimedia presentation might be more enjoyable to work with compared to for example problem lists because it presents information in an engaging and varied manner. The multimedia presentation took approximately three days to prepare.

21

Kort beskrivelse: Flere brugere overser søgeresultaterne og påbegynder i stedet ny søgning.

Alvor:

Level 3. Det bliver opfattet som om den første søgning (fra 'google-siden') bliver ignoreret af systemet. Der kan opstå tvivl om, hvorvidt systemet rent faktisk registrerer/handler på de ting, som brugeren indtaster.

Figure 3: Example of problem list (PRO) format.

The redesign proposals (RED, Figure 6) each consisted of a brief description of the usability problem, a description of one or more solutions, a justification of the solutions, illustrations of the solutions, and finally a short text explaining the illustrations. Redesign proposals were included because justifications should make them convincing and because the ideas for solutions should improve the understanding of the usability problem, and facilitate the actual fixing of a problem. The redesign proposals took approximately a day and a half to prepare.

Representing scenarios (SCE, Figure 7) we chose to use human-centered stories. These were expected to be persuasive and to provide valuable information about the context of use. In this study a scenario was approximately one page long, included six lines of introduction (presenting the characters and 'setting the stage'), and a narrative that described a problem, the context, and the user's motivation and feelings in the situation. The scenarios took approximately two days to prepare.

The feedback was presented to the developers on paper (formats PRO, SCR, RED, SCE) and CD-rom (format MUL). We found this most flexible and according to practice. Despite numerous recommendations to interact with developers (Kahn & Prail, 1993; Butler & Ehrlich, 1993; Dumas & Redish, 1999), this study refrained from studying oral feedback. This was done to emphasize the importance of the deliverables that support oral feedback and that serve as documentation and reminders for developers during their work.

Producing comparable feedback items

The five formats PRO, SCR, MUL, RED, and SCE consisted of a combination of descriptive elements such as text, illustrations and severity ratings. We produced a series of descriptive elements to be copy-pasted when we constructed the feedback according to the five formats. We did this to improve the comparability between the formats. For example, the same rating would be used for all formats presenting a severity rating of a specific problem. Step four resulted in a total of 200 so-called feedback items, comprising 40 usability problems described by five feedback formats (see Figure 1). Examples of all formats can be found in the appendix.

Step five—Pre-use rating of feedback items

To rate the value of the feedback items, the 200 items were presented to three developers at Jobindex who usually receive and take care of usability feedback. A description of the test set-up, the participants, and the tasks were also provided.

The 200 items were presented in random order so that no one feedback format was favoured by being presented first. Each feedback item was presented with a rating sheet where each developer individually would assess every feedback item according to the questions in Table 1. The questions were intended to shed light on issues such as usefulness, persuasive power, and clarity; issues that are crucial for the feedback's quality. These questions also aimed to make sure that formats were assessed independently of how difficult each usability problem was. To answer the questions, the developer would mark a point on a 100 mm horizontal line. Each end of the line was marked with the labels shown in parenthesis after the questions (e.g., 'very poorly'/'very well', see Table 1). This method of measuring has been used in other studies (Hornbæk & Frøkjær, 2005) and lets the developers answer questions without being constrained by a small number of categories on the scale. The scale was quantified by measuring the millimetres from the start point to the point on the line marked by the developer. Each developer used approximately four hours rating the feedback items.

Step six—putting the feedback items into action

After developers had rated their first impressions of the feedback, we wanted to study how they would use the feedback in their daily work. Each developer received a set of the 40 usability problems; 32 problems in print (covering equally feedback formats PRO, SCR, RED, SCE), and the remaining eight problems on a CD-rom (MUL). The feedback items were selected at random from the set of 200 feedback items produced in step four, so that the developers would only work with each of the 40 problems once.

The developers were instructed to carry out their work on the system as if they had received any other usability report. This was done so the developers could familiarize themselves with, and perhaps change their opinions about the feedback items once they got experienced using them. The developers worked with the feedback items for approximately 12 weeks together their other

tasks at Jobindex, as they would have done with the feedback they usually receive.

Step seven—post-use rating of feedback items

Having finished the work on the system, the developers repeated step five, this time rating only the 40 feedback items they had been working with, keeping their actual work experiences in mind. Each developer used approximately one and a half hours on this task.

Step eight—individual interviews.

Finally, developers were interviewed individually. They were presented with and asked to discuss examples of the feedback items which they had rated the highest and the lowest. The developers were also asked to discuss the significance of positive findings and to perform a card sort during which they discussed the value of specific feedback elements such as severity ratings, video, and contextual screen dumps. The aim of the interviews was to collect finer nuances on the developers' opinions, anecdotal data about their experiences with the feedback formats, and ideas for improving feedback on usability evaluation. During the interviews, the developers' opinions were noted directly on the relevant feedback items. The interviews were afterwards documented with thorough notes, two of the interviews were additionally audio recorded.

Results

Pre-use ratings

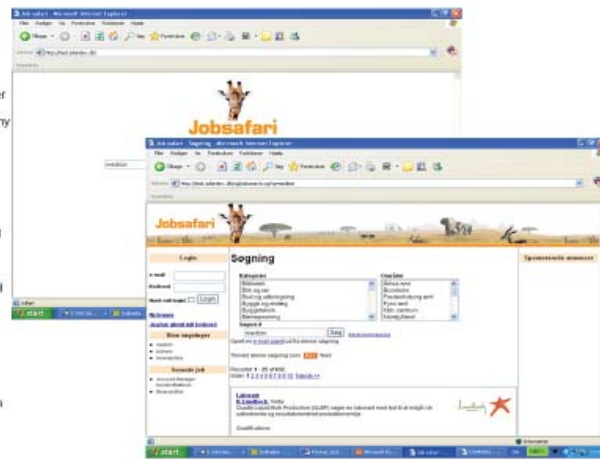
Table 2 presents an overview of developers' mean pre-use ratings of the five feedback formats. To protect the experiment-wide error, we first analyzed the pre-use ratings using multivariate analysis of variance (MANOVA). This test suggests significant differences between how feedback formats perform.

Post hoc tests show that redesign proposals, the multimedia presentation, and the screen dump format were rated equal. They were rated significantly better than the problem list, which in turn was rated better than scenarios. As an example,

21

Kortbeskrivelse:
Flere brugere overser søgeresultaterne og påbegynder i stedet ny søgning.

Løsningsforslag:
Boksene med rullemenuer på resultatsiden tager al opmærksomheden. De kan måske erstattes af et enkelt firkantsfelt og et link til en avanceret søgning. Det vil give mere plads at vise resultater på. Dermed tilbyder resultatsiden de samme søgemuligheder som 'google-siden', der kan fjernes.



Brugeren opdager ikke søgeresultaterne når han klikker søg på 'google-siden' og kommer frem til resultatsiden.

Figure 4: Example of screendump list (SCR) format.

developers rated redesign proposals highest in 40% of the cases, the multimedia presentation in 31% of the cases, the screen dump format in 23% of the cases, and the problem list in 6% of the cases. Scenarios were never rated highest.

To investigate these differences, we conducted individual analyses of variance on each question. Table 2 shows how the significant groups changed among questions. Among the three top-rated formats (screen dumps, multimedia presentation and redesign proposals), the redesign proposal format is rated significantly higher than the multimedia presentation on a question concerning whether a feedback item helps the developer solve the problem. The multimedia presentation seems to be slightly better at convincing the developer about the problem, but the difference to the screen dump format and the redesign proposal format is not significant. Despite the small variances on individual questions the ratings generally support the picture from Table 1; the screen dump format, the multimedia presentation and redesign proposals were the most valued feedback formats.

We found no significant effect of order of presentation or ratings, $F(7,167) = 0.54, p < .921$, suggesting that having seen a usability problem presented by one or more formats did not affect how a developer rated a feedback item.

Post-use ratings

Table 2 also shows the mean post-use ratings. An overall MANOVA test showed no significant difference in how the five formats were rated after use. An analysis of the ratings of each individual ques-

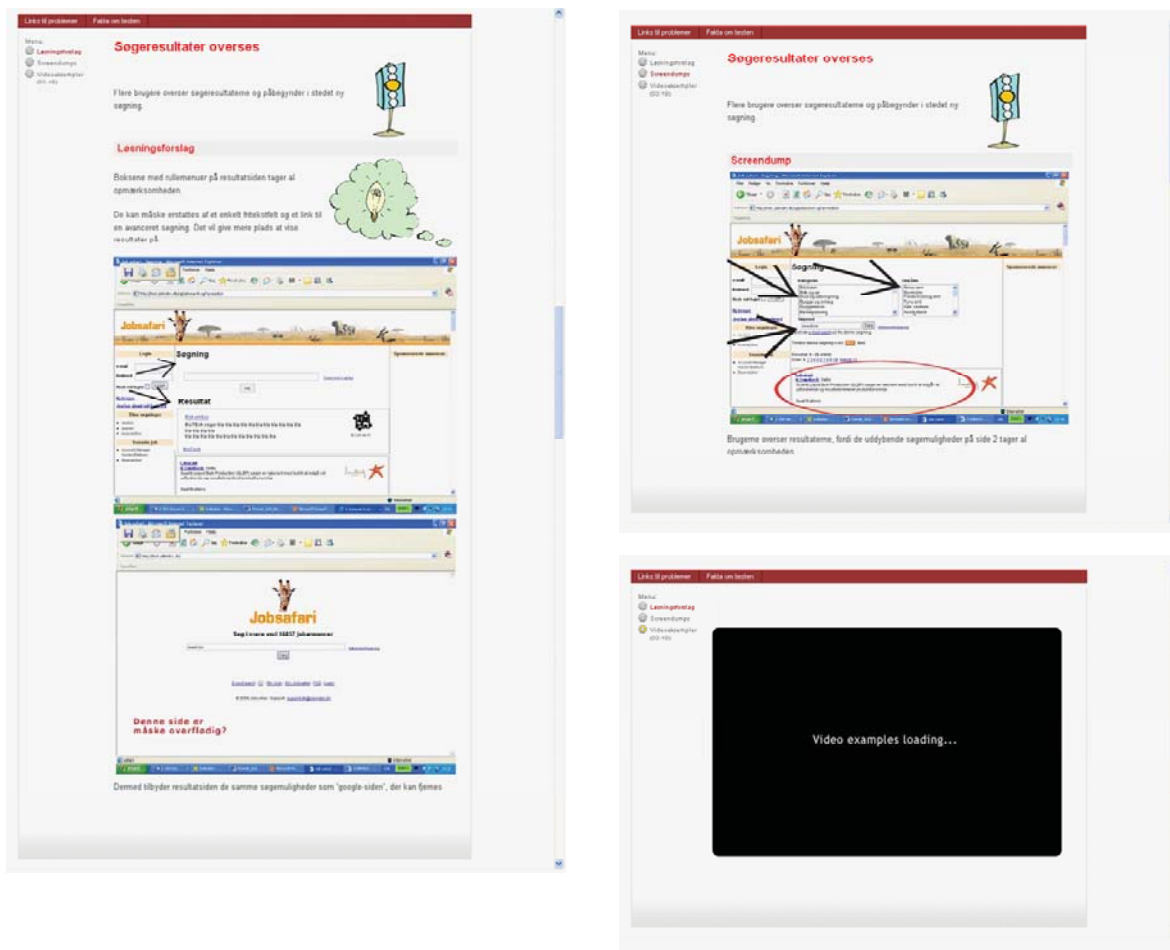


Figure 5: Example of multimedia list (MUL) format.

tion confirmed this result.

A comparison of the ratings of identical feedback items pre-use and post-use (Figure 8) showed that all five questions received lower post-use ratings. The only exception was scenarios (SCE), which generally received the same rating.

Interviews

We consolidated the notes from the interviews into 14 groups of similar opinions. Four of these identified general parameters that make feedback useful to developers, namely that the problem can be recognized, that the problem is easy to fix, that the feedback contains much information about the problem’s context, and that the feedback is quick and easy to use. Table 3 provides examples of the developers’ opinions regarding the strengths and weaknesses of the feedback formats.

General findings – explaining high and low ratings

The interviews showed that the top rated feedback items shared some characteristics. First,

the problems were recognizable to the developer, meaning that the developer knew about them already. As an example, developer 3 (Dev3) explained:

This is a much more recognizable problem. I know it is annoying. It is a problem I have been in contact with before’. Second, five of the ten highest rated feedback items were considered easy to fix: ‘It’s a change that can be easily overcome...that’s why it has a higher rating (Dev3).

Six of the ten highest rated feedback items were rated highly because the developers agreed with the problem.

The lowest rated feedback items also showed similarities. A feedback item that received a low rating often described a problem that was hard to recognize, either because the developer was not convinced about the problem or because he needed more contextual information to understand it. Dev2 pointed to one reason for not being convinced about the relevance of the problem and wanting to know more of its context: ‘I am not

able to deduce the cause of the problem from this feedback'. Developers explained that for five of the ten lowest rated feedback items they disagreed with the problem or found it impossible to solve. Developers explained four of the ten lowest rated items with not being able to understand the problem, for example:

I have trouble understanding what it is...I mean what search words the user typed...I understand that the user has typed something and has an expectation about finding something...but I have a hard time understanding what it is (Dev1).

Generally, developers valued the access to contextual information, and several formats were criticized for not describing enough context: 'I need to know more', Dev1 pointed out during the discussion of the lowest rated feedback items. Conversely, formats with rich descriptions of context are seemingly not without problems. Feedback formats that elaborated on context of use were either criticized for being tedious to use (multimedia presentation) or rated poorly throughout the study (scenarios). This suggests that developers consider a format's ease of use an important parameter when assessing how a format performs.

Details on the five formats

Developers criticized scenarios for being time consuming and 'full of noise' (Dev1). Dev3 argued that the scenario format did not really help to fix the problem and that it was often difficult to understand a problem. 'It does surprise me that I can still be unsure of what the problem is after having read this long text', he explained. On the positive side, the scenario format 'shows you where you lose the user' (Dev1) and provides contextual information about the usability problem, which help understanding what the evaluator had in mind (Dev2).

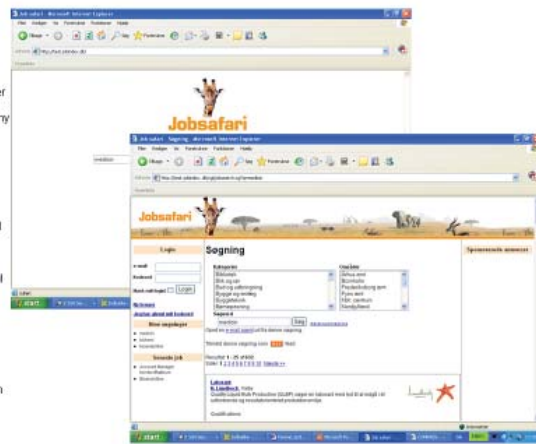
Problem lists were considered suitable for presenting uncontroversial usability problems. Dev2 described how he used the severity rating to estimate whether he was 'on the same level as the evaluator' and that agreeing with the severity ratings meant that he perceived findings as more valid. Dev3 criticized the problem list format for lacking contextual information:

The problem has been boiled down to one

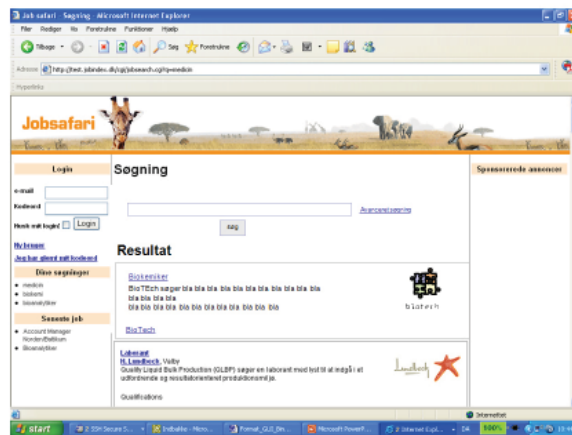
21

Kort beskrivelse:
Flere brugere overser søgeresultaterne og påbegynder i stedet ny søgning.

Løsningsforslag:
Boksene med rullemenuer på resultatsiden tager al opmærksomheden. De kan måske erstattes af et enkelt linksfelt og et link til en avanceret søgning. Det vil give mere plads til vise resultater på. Dermed tilbyder resultatsiden de samme søgemuligheder som google-sider, der kan fjernes



Brugeren opdager ikke søgeresultaterne når han klikker søg på 'google-siden' og kommer frem til resultatsiden.



Fjernes katrongsøgningen fra resultatsiden, trækker resultatet af søgningen tydeligere frem.

Figure 6: Example of redesign proposal (RED) format.

line of text. It can be difficult to understand [the problem] because the description is too short and it does not include any description of context, suggestions for solutions or anything. I often catch myself thinking what am I supposed to do with this?

Dev2 valued the multimedia presentation in particular for the videos. He explained how videos provide fine nuances about the context and the use of the system. Dev1 and Dev3 valued the possibility of exploring a video, but found that the usability problems were generally easily understood without seeing the video. Describing simple usability problems with video seemed unnecessary to them, and they criticized the multimedia presentation for being too time consuming to use because of the videos. Dev1 explained how he found the video in the multimedia presentation tedious because it was difficult to get a quick overview and to skim the content: 'It doesn't allow me to fast forward to exactly where the problem occurs'. He suggested providing a textual description of the video's story line, using scenarios as a model.

21

Historien om Maria og Frank

Maria er højgravid og fylder godt ud i det snævre køkken i lejligheden på Vesterbro. Frank ynder at lave sjov med, at de bliver nødt til at rive væggen til gangen ned, hvis hun skulle finde på at starte sine veer foran komuret. Maria ynder at minde Frank om, at der faktisk er 10 dage til termin, og at hun kan nå at blive meget store. Frank og Maria læser bokemi, han er netop begyndt at skrive sit speciale og hun er snart færdig med sin ph.d. Nu står en familieførelse for døren og parret overvejer om Frank skal prøve at få noget arbejde, for at supplere deres økonomi. De har ikke mange penge mellem hænderne for Frank får ikke længere SU, fordi han tidligere har studeret medicin.

Brugeren opdager ikke søgeresultaterne

Solen skinner ind ad Marias vindue på Panum. Han sidder og surfer efter jobs, mens hun småsnakker med Lone, som hun deler kontor med.
"Finder du noget?", spørger Lone.
"Tja, det ser ikke helt skidt ud", siger Maria.
"Kom over og se".

Lone rejser sig fra stolen og går over og stiller sig ved siden af Maria.
"Prøv at se, hvordan det ser ud i København for forskere", siger Lone og læner sig ind over tastaturet.
"Må jeg?" spørger hun, og indtaster 'bokemi' og forsker i søgefeltet.

Lone kommer ind på listen over resultater.
"Hmmm..." hun rynker lidt på næsen.
"Det her er så en ny søgeside", siger hun løvende.
Hun vælger emnet 'medicinal og biotech' fra rullemenuen over emner og 'København' fra rullemenuen over amter. Så klikker hun 'søg' igen.

"Det er alligevel lidt åndssvagt. Så du, hvordan den bare ignorerede hvad jeg skrev?", siger Lone.
"Jeg tror nu at du overså resultaterne far", siger Maria, og klikker på Browserens 'tilbage-knap'.
"De står her lidt nede på siden. Under de store fede bogstaver, der siger 'Resultater'.
"Ja-ja, det er godt med dig, Karl Smart", siger Lone, og tager musen fra Maria.

Figure 7: Example of scenario (SCE) format.

Dev3 supported this idea. Further, the developers did not find that graphical illustrations added any value and called for more thoroughly explained severity ratings.

Dev1 commented that screen dumps were easier to understand than plain text which is often imprecise. Dev2 repeated this point for the textual redesign proposals; text can be difficult to understand, thus illustrations of redesign proposals are called for.

Dev2 explained how the redesign proposals in the redesign proposal format made it easier to understand and accept critique. He explained how the evaluator's efforts to illustrate a redesign idea improved the quality of these ideas. All three developers agreed that the justification for the redesign proposal is unnecessary: 'A good idea should speak for itself', said Dev1.

The feature of directly pointing to where the usability problem occurred received positive comments from all three developers. Dev2 explained how the multimedia presentation let him navigate from the problem description to an illustration of where the problem occurred, making the multimedia presentation easy to use. Dev3 pointed out a positive feature of the screen dump format: 'It gets pinpointed where it [the problem] is'. The redesign proposals, the screen dump and the multimedia presentation formats all included the feature of illustrating where the usability problem occurred.

Characterization of the usability problems

The ratings of different feedback formats may depend on the nature of the problems that are used

in this study. To investigate this, five researchers rated the 35 usability problems according to (a) discoverability; how easily they were discovered, and (b) complexity; the perceived complexity of fixing the problem. Discoverability was coded according to the scale perceivable, actionable and constructable (Cockton & Woolrych, 2001). Perceivable problems are the easiest to discover and can usually be discovered by simply looking at the display. Actionable problems can be identified with one to a few steps or clicks. Constructable problems are the hardest to identify and need several steps of interaction to be revealed. Complexity was coded following Hornbæk and Frøkjær (2004) using a three-step scale

comprising complex, medium-sized and simple problems. The average complexity-discoverability ratio is shown in Table 4 and suggests that the usability problems used in this study were mostly simple and perceivable/actionable.

To get an impression of whether the most heavy-weight usability problems were rated differently than the rest of the usability problems, we studied the ratings of the six usability problems from the bottom-right corner of Table 4 (being the actionable-constructable/ middle-complex problems). On average, the heavyweight usability problems were rated 8% lower than the more light-weight usability problems pre-use, though this difference was not significant, $F(1,173) = 1.757, p > .1$.

Low answering rate for positive findings

On average, developers answered 95% of the questions pre-use and 98.5% post-use. The only apparent pattern among the unanswered questions was a low answering rate for positive findings. This is unsurprising since three of five questions specifically concerned usability problems, and was irrelevant for positive findings. However, during the interviews the developers expressed general satisfaction with receiving positive findings and explained that it was nice to know which parts of the system that worked well and should not be changed. Dev2 also mentioned the psychological effect of combining negative with positive feedback for the critique to be 'easier to swallow'.

Discussion

Comparing feedback formats

Our explorative study suggests that the multimedia presentation, the screen dump format, and the redesign proposals were generally seen as useful input to developers' work whereas the scenarios were not well received. The problem list was generally rated lower than the three top formats but higher than the scenarios.

The study suggests that feedback serves several functions, the relative importance of which change over time. Understanding the problem and being convinced about it seems of initial importance. Information about a problem's context apparently plays a role for how well a problem is understood and for how convincing developers find it. Contextual information seems to elaborate on the problem, making it easier to understand, and provides information on what caused the problem thus making it more convincing. When the developer is convinced about the problem and understands it, it becomes important whether the feedback is easy to use. Ease of use and thorough contextual information seem quickly to conflict however, since the adding of more contextual data seems to make feedback more time consuming to use. Finally, when the developer has worked with a problem for a while the feedback mainly serves as a reminder about the problem. Below we discuss how the five feedback formats support these functions.

The problem list was generally rated lower than the screen dump format, the multimedia presentation, and redesign proposals, suggesting that the most commonly used feedback format is not the most effective one. Problem lists seem best suited for communicating simple and uncontroversial usability problems for which no contextual information is needed. We argue that some of the recommendations to improve problem lists such as 'be more positive, clear, precise and respectful' (Dumas, 1989) do not fully address the challenges associated with problem lists. Problem lists do not provide any explanations to bolster its problem description, and the format's ability to convince seems mostly to rest on the evaluator's ethos and assertiveness, as argued by Nørgaard and Høegh (2008). Nørgaard and Høegh showed that developers used severity ratings to assess the evaluator's credibility and concluded that well argued severity ratings make problem lists more credible.

		Pre-use	Post-use
Problem List	M SD	45.3 12.5	40.5 14.0
Screen dump	M SD	54.0 10.9	42.1 17.1
Multimedia presentation	M SD	53.8 12.1	50.2 14.6
Redesign proposal	M SD	57.0 11.2	43.5 12.6
Scenario	M SD	31.6 9.9	32.4 14.0
F-test		$F(4,170)=28.76, p<.001$	$F(4,30)=1.35, p<.3$
Tukey HSD post hoc test		SCE<PRO< MUL,SCR,RED	SCE,PRO, MUL,SCR,RED

Table 2: Average pre-use and post-use ratings of questions 1-5 for each feedback format. The rating spans the numbers 1-100, 100 being the best rating.

The screen dump format, which can be produced at fairly low cost, was generally rated similarly to the multimedia presentation and redesign proposals. The context provided by the annotated screen dumps was valued greatly by developers as conveying a better understanding of the problem. The screen dump format only showed where the problem occurred and gave no information about what led to the problem as did the multimedia presentation. This difference in content compared with the difference in how the two formats were rated suggests that information about problem occurrence is more important to developers than contextual feedback about for instance users' interactions with the system.

The multimedia presentation proved less convincing than suggested by the literature on highlights videos. 'Seeing is believing' is a common argument for videos (Desurvire, Lawrence, & Atwood, 1991), but our study suggests that other formats are equally convincing. Developers called for an easier access to contextual information than video. This critique points to contextual information, like the one presented by the videos in the multimedia presentation, as being important to understanding usability problems.

The high ratings of redesign proposals suggested that they served as a valuable elaboration of the problem description and made the usability problems more understandable to developers.

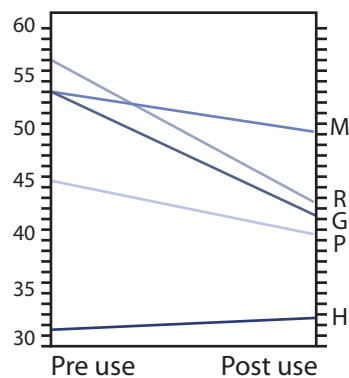


Figure 8: The ratings of formats pre-use and post-use.

This finding supports the findings of Hornbæk and Frøkjær (2005), and suggests that the quality of feedback on even fairly simple problems is improved by the use of redesign proposals. The psychological effect of receiving constructive suggestions rather than negative criticism may also partially explain why developers found the format convincing.

Scenarios performed poorly on the question regarding how well a problem was understood. One reason may be that they demanded the reader to analyze and interpret the scenario before being able to understand the problem. The fictional style of the presentation might also have been problematic since developers found it unconvincing. This problem could perhaps be addressed by modifying the narrative style of writing. However, we need to stress that the type of scenario used in this study, the human-centered story, was not designed for providing feedback on usability problems, and we suggest investigating how other types of scenarios can be used to improve feedback from evaluations.

Feedback issues of importance to developers

Our study suggests that developers rate usability problems with which they agree higher than the ones with which they do not agree. This finding underlines the importance of feedback formats' ability to convince. Problems that are easily fixed seems also to be rated higher than problems that are not easily fixed. This finding supports reports on how developers often favour the problems which are easiest to correct (Dumas & Redish, 1999).

Developers value contextual information, which may explain why the multimedia presentation, screen dump format, and redesign proposals, which all described context such as problem oc-

currence, were initially preferred by developers. We speculate that the need for contextual information is linked to developers' wish to investigate certain usability problems in depth to obtain a better understanding of the problem or to search for convincing factors about the problem.

Differences in pre-use and post-use ratings

The differences in how developers rate feedback formats diminish after they have worked with the feedback items. Since we expected the developers to familiarize themselves with and develop preferences for certain formats during their work, we were surprised to find that the post-use rating showed no significant differences among the formats.

Even though developers answered the same questions before and after having worked with the problems, we hypothesize that the questions were perceived differently pre-use and post-use. As we have no way of knowing, we will refrain from speculating what the difference in meaning is. We dare to speculate that when learning about and having to understand a problem, annotated screen dumps, multimedia presentations, and redesign proposals are superior to problem lists and scenarios. Yet, any of the five formats may serve as a reminder of a specific problem. This study does not present enough evidence to finally confirm such speculations, and we suggest conducting more studies to identify which feedback formats that improve downstream utility.

Recommendations

Developers seem sensitive to information overload, and we need to investigate how thorough contextual information can be presented in the least overwhelming manner. A multimedia presentation that allows developers to study relevant information and discard irrelevant information might be a solution. Indexed videos might speed up navigation in a short highlights video, but some problems are difficult to explain in a short video, because they show over several minutes of use and cannot be presented in a few frames. Written summaries of what happens in a video clip might make contextual information more accessible. Such summaries could resemble scenarios and leave room for the evaluator to elaborate on problems that are difficult to illustrate with short video clips. However, since scenarios are not considered a convincing format this idea is not without problems. Longer problem descriptions with elaboration of the causes of the problem might

	Strengths	Weaknesses
Problem List	Provides short and sufficient information about simple problems. Ratings of severity.	Does not describe context of problem. Too short to describe problems fully.
Screen dump	Points to where the problems should be fixed. Concrete and often easier to understand than text.	The problem's context and triggers need to be elaborated. An illustration of the redesign proposal is lacking.
Multimedia presentation	Video is credible and persuasive. Quick and easy to use.	'Overkill' to describe simple problems with video. Video is too time consuming and it is difficult to get a quick overview of the video.
Redesign proposal	Helps solve the problem well. Illustrations improve quality of redesign proposals.	The problem's context and triggers are not explained well. A justification is unnecessary.
Scenario	Provides information about the context of a problem. Shows where you 'loose' the user in the interaction.	'Overkill' - it is not a simple way to present a problem. There is a lot of 'noise'. Time consuming to read and interpret.

Table 3: A description of the feedback formats' strengths and weaknesses

also improve problem lists. Screen dumps that show where a problem occurred seem also to be valued information that is easily produced, and may even serve as reference for a redesign proposal.

Future Work

Since this study is explorative it does not draw on a large collection of data derived from many types of companies, numerous subjects or different use contexts. We recommend that our assumptions be validated by further studies. Such studies could investigate more informants, different types of companies and different ways of working with usability feedback.

The questions used to obtain the ratings in this study also deserve attention. We compared how the ratings of the five questions varied, and the results showed that the questions did not receive significantly different ratings. This suggests that the developers might have had difficulty understanding the nuances of the five questions. Though the questions were formulated to help developers rate specific aspects of a feedback format, it is also possible that the severity of a problem influenced how it was rated. Future studies should pay

more attention to how questions are formulated, perhaps using examples to illustrate the aim of a question. Also, studies that include questions regarding how problem severity is perceived might help us understand more precisely if the severity of a problem has something to do with how a format is valued, or if the same problem described by different formats is perceived as having differing severity.

The scenario format received poor ratings in this study. However, before rejecting scenarios as a way to deliver important information about user experience and context, we recommend that a broader range of scenario-methods be studied to, for example, investigate if scenarios can be used to provide certain contextual information that is perhaps difficult to describe with highlights videos.

This study suggests that the role of feedback changes over time. Accordingly, studies that solely concern pre-use evaluation results are problematic because they miss the post-use aspects of feedback. To address this problem we call for future work on the various stages and roles of feedback such as how different types of feedback

	Perceivable	Actionable	Constructable
Simple	11	15	0
Middle	3	4	1
Complex	0	1	0

Table 4: The 35 usability problems sorted according to complexity and discoverability.

are used in different phases of the development work.

Conclusion

This study investigated how five feedback formats served to convince developers of the existence of the problems and to convey an understanding of the usability problems. The study suggests that feedback serves multiple purposes which change during its use. Initially, feedback needs to convince developers that problems exist and to help them understand the problems. The amount of contextual information is crucial to how well a feedback format succeeds in convincing developers about the relevance of a problem. Having accomplished that, feedback must be easy to use in the developers' daily work. Thereafter it mainly serves as a reminder of a problem.

These findings point to a problem in earlier studies that seem to understand feedback as a static product. As an example, studies that look only at first impressions of feedback may come to very different conclusions about the quality of a format than studies that at similar format after it has been used in a work situation. To fully understand the implications of this study we first need to validate the findings by conducting more studies that investigate feedback before and after use. Such studies must examine the roles of feedback in various work situations and in various organizational contexts.

Specifically regarding the five formats studied in this paper, developers rated the multimedia presentation, redesign proposals and the screen dump format highest on first hand impression. After having worked with the feedback, developers rated problem lists, the screen dump format, the multimedia presentation, redesign proposals, and the scenario format alike. The findings suggest that all feedback formats may serve as a reminder, but that only some formats convey the

information needed to initially portray a problem clearly and convincingly. The problem lists used in this study did not provide sufficient information to perform well on first hand impressions. This suggests that this commonly used feedback format needs to include additional information to provide developers with efficient feedback.

References

- Butler, M., & Ehrlich, K. (1993). Case study: Lotus Notes 1-2-3 Release 4. [http://domino.watson.ibm.com/cambridge/research.nsf/0/1b6720e0274ed7b3852563bf0062325f/\\$FILE/Wiklund.pdf](http://domino.watson.ibm.com/cambridge/research.nsf/0/1b6720e0274ed7b3852563bf0062325f/$FILE/Wiklund.pdf).
- Coble, J., Karat, J., & Kahn, M. (1997). Maintaining a focus on user requirements throughout the development of clinical workstation software. *Proceedings of the ACM Conference on Human Factors in Computing*, 170-177.
- Cockton, G., & Woolrych, A. (2001). Understanding inspection methods: lessons from an assessment of heuristic evaluation. In A. Blandford, J. Vanderdonck, & P. Gray, *People and computers XV - Interaction without frontiers. Joint Proceedings of HCI2001 and IHM2001*, 171-191, Springer.
- Desurvire, H., Lawrence, D., & Atwood, M. (1991). Empiricism versus judgement: Comparing user interface evaluation methods on a new telephone-based interface. *ACM SIGCHI Bulletin*, 23, 4, 58-59.
- Dumas, J. (1989). Stimulating change through usability testing. *SIGCHI Bulletin*, July 1989, 21, 1, 37-44.
- Dumas, J., & Redish, J. (1999). *A practical guide to usability testing*. Oregon, USA: Intellect Books.
- Dumas, J., Molich, R., & Jeffries, R. (2004). Business: Describing usability problems: Are we sending the right message? *Interactions*, 11, 4, 24-29.
- Hornbæk, K., & Frøkjær, E. (2005). Comparing usability problems and redesign proposals as input to practical systems development. *ACM Conference on Human Factors in Computing Systems*, 391-400.
- Hornbæk, K., & Stage, J. (2006). Special issue on the interplay between usability evaluation and user interaction design. *International Journal of Human-Computer Interaction*, 21, 5.
- Jeffries, R. (1993). Usability problem reports: Helping evaluators communicate effectively with developers. In J. Nielsen, & M. R.L., *Usability inspection methods*, 273-294, John Wiley & Sons.
- Jeffries, R., Miller, J., Wharton, C., & Uyeda, K. (1991). User interface evaluation in the real world: A

- comparison of four techniques. ACM Conference on Human Factors in Computing Systems, 119-124.
- John, B. E., & Marks, S. J. (1997). Tracking the Effectiveness of Usability Evaluation Methods. *Behaviour & Information Technology*, 16, 188-202.
- Kahn, M., & Prail, A. (1993). Formal usability inspections. In J. Nielsen, & R. Mack, *Usability inspection methods*, 141-171, John Wiley & Sons.
- Kennedy, S. (1989). Using video in the BNR usability lab. *SIGCHI Bulletin*, 21, 2, 92-95.
- Law, E. (2006). Evaluating the downstream utility of user tests and examining the developer effect: A case study. *International Journal of Human-Computer Interaction*, 21, 2, 147-172.
- Mills, C. (1987). Usability testing in the real world. *SIGCHI Bulletin*, 18, 67-70.
- Nayak, N., Mrazek, D., & Smith, D. (1995). Analyzing and communicating usability data. *SIGCHI Bulletin*, 27, 1, 22-30.
- Nielsen, J. (1993). Heuristic evaluation. In J. Nielsen, & R. Mack, *Usability inspection methods*, 25-62, John Wiley & Sons.
- Nørgaard, M., & Høegh, R. T. (2008). Evaluating Usability - Using Rhetorical Models to Improve the Persuasiveness of Usability Feedback. Proceedings of the 7th ACM Conference on Designing Interactive Systems (DIS2008) .
- Redish, J., Bias, R., Bailey, R., Molich, R., Dumas, R., & Spool, J. (2002). Usability in practice: Formative usability evaluations - Evolution and revolution. ACM Conference on Human Factors in Computing System, Minneapolis, Minnesota, 885-890.
- Rosson, M., & Carroll, J. (2002). Usability engineering: Scenario-based development of human-computer interaction. Morgan Kaufman.
- Schell, D. (1986). Usability testing of screen design: Beyond standards, principles, and guidelines. Proceedings of the Human Factors Society 30th Meeting, Santa Monica, CA, 1212-1215.
- Sears, A. (1997). Heuristic walkthroughs: Finding the problem without the noise. *International Journal of Human-Computer Interaction*, 9, 3, 213-234.
- Seffah, A., & Andreevskaia, A. (2003). Empowering software engineers in human-computered design. Proceedings of the 25th International Conference on Software Engineering, 653.
- Strøm, G. (2003). Perception of human-centered stories and technical descriptions when analyzing and negotiating requirements. Proceedings of Human-Computer Interaction, Interact '03, 912-915.

Evaluating Usability

Using Models of Argumentation to Improve Persuasiveness of Usability Feedback⁶

Mie Nørgaard

Computer Science
University of Copenhagen
mien@diku.dk

Rune T. Høegh

Computer Science
Aalborg University
runethh@cs.aau.dk

Abstract

Usability evaluation is widely accepted as a valuable activity in software development. However, how results effectively are fed back to developers is still a relatively unexplored area. We argue that usability feedback can be understood as an argument for a series of usability problems, and that basic concepts from argumentation theory can help us understand how to create persuasive feedback. We revisit two field studies on usability feedback to study if concepts from Toulmin's model for argumentation and Aristotle's modes of persuasion can explain why some feedback formats outperform others. We recommend that evaluators specifically back up the warrants behind their usability claims, that their arguments use several modes of persuasion, and that they present feedback in browsable amounts not to overwhelm developers with information. For complex and controversial problems, we advise evaluators to involve developers in a learning process and provide the opportunity to experience and discuss the findings.

⁶ This paper was originally published in Proceedings on the 7th ACM conference on Designing Interactive Systems (DIS'08), February 25th-27th, 2008, Cape Town, South Africa.

Introduction

Imagine receiving the following comment on a system, you have been working on for several months: ‘The user cannot find what he is looking for’. Would such a description make a big impression on you? Or would you somehow not be convinced about the nature of the problem? Probably the latter.

With respect to usability evaluations, the challenge of feedback is to move beyond solely presenting descriptions of usability problems (UPs) to actually make the receiver understand and acknowledge the nature and relevance of the problems. The example above does the first but probably not the latter.

When working with usability we see at least two different approaches to improving the impact of usability results: one approach, which focuses on improving usability evaluation methods (UEMs) and practices, and another approach, which focuses on how evaluation results are communicated and fed back into the development process. These approaches are fundamentally different. The first approach builds on the perception that UEMs somehow do not provide sufficiently useful or relevant usability results, and that the matter of impact should be addressed through improved UEMs. The other approach builds on the perception that, while UEMs might still be improved, the real bottleneck is having the developers acknowledge the identified problems.

In this article we set out to investigate how to strengthen the impact of usability by improving the quality of the feedback. But while previous work on usability feedback seems to understand feedback as merely a presentation or description of problems, see for instance (American National Standards Institute, 2001; Dumas & Redish, 1993; Mills, 1987; Redish, Bias, Bailey, Molich, & Spool, 2002; Rubin, 1994), we aim to investigate whether it makes sense to view usability feedback as an argumentation that seeks to make developers acknowledge the nature and relevance of usability problems. We hypothesize that understanding feedback in this frame will help us explain why some pieces of feedback are more persuasive than others.

Before progressing any further, a few words on feedback from usability evaluations are needed. We recognize that physical deliverables, such as a usability report, are often presented as part of a feedback process, where usability evaluators orally present findings, elaborate on results, or offer developers the opportunity to discuss the findings. However, in this study we view the de-

liverables isolated from the delivery and work context. We do so because we believe that deliverables have a certain value for example as input to discussions, or as reminder for the developers, who are working on the system. Accordingly, we believe that the deliverables could be studied isolated from the feedback process in order to better understand what a useful deliverable is. However, our study includes two examples of feedback formats where written deliverables play only a minor role. In these cases the process of experiencing or discussing UPs is the main contributor of feedback. We include these examples of feedback because comparing written feedback to learning-oriented feedback might help us uncover strengths and weaknesses of written feedback. Also, we expect to learn something about, not only the role of arguments in feedback, but also about important pedagogical aspects of feedback such as the value of having developers experience usability problems for themselves.

We re-examine two field studies to investigate if the success and failure of feedback formats can be understood in terms of argumentation theory. We identify how formats that present usability arguments in a way that corresponds with concepts from Toulmin’s model for argumentation and Aristotle’s three modes of persuasion seem more persuasive than the ones that do not. Finally, we discuss the implication of these findings, and recommend that producers of feedback make an effort to identify and back up the warrants behind their usability claims. We further recommend that arguments use several modes of persuasion, and that they present feedback in browsable amounts in order not to overwhelm developers with information. For controversial UPs we recommend feeding back information in manners that resemble the two learning-oriented formats.

Related work

In the past, the predominant feedback format has been the usability report. Classical usability literature, such as (American National Standards Institute, 2001; Dumas & Redish, 1993; Mills, 1987; Redish, Bias, Bailey, Molich, & Spool, 2002; Rubin, 1994), recommend using the report format, and present advice on how to put together the content. Some advice relates to the choice of words such as avoiding technical jargon, including positive findings and expressing problems tactfully (Dumas & Redish, 1993). To demystify usability work Molich (2000) suggests that usability practitioners invite developers and other receivers of feedback to see the test facilities and get a demonstration of the methods in use.

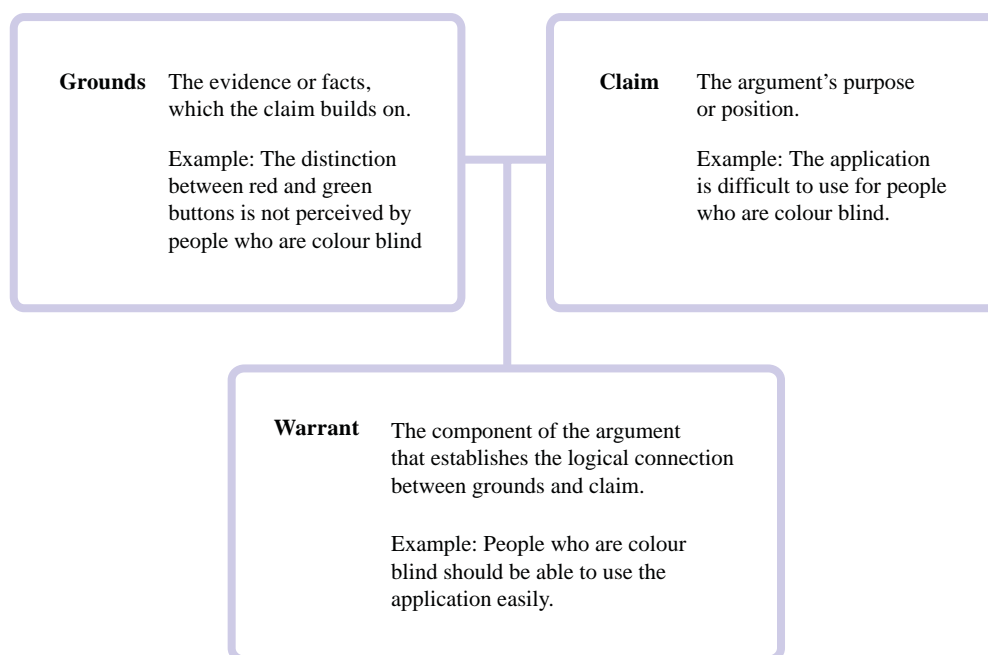


Figure 1: An example of how a usability claim might relate to grounds and warrants. The grounds provide the evidence for the claim, and the warrant describe the underlying (and often implicit) assumptions that must be agreed upon before the claim can be accepted. Adapted model after (Toulmin, 1958).

A recent literature study (Høegh, 2006) concluded that little research presents a critical view on how results are fed back from usability evaluations to the development process. Studies concerned with feedback agree that designing feedback to make an impact is a complicated process, see for instance (Boivie, Åborg, Persson, & Löfberg, 2003). John and Marks have studied the effectiveness of UEMs, by tracking their impact on software (John & Marks, 1997). They experienced that only 50% of the reported problems led to changes in the code, and that UEMs are not as effective, as most practitioners would like to think. Law (2006) discusses and defines the concept developer effect as developers' systemic biases to usability problems with particular characteristics. She argues that such biases will heavily influence the number of usability recommendations being implemented in a system. In order to improve what she calls downstream utility Law points to the importance of successfully convincing developers that the feedback is relevant.

Rhetoric and argumentation

Since we understand feedback from usability evaluations as an argument for usability problems, we believe that knowing concepts from classical rhetoric and argumentation theory is valuable

for usability practitioners. Next, we present two views on argumentation that can help us to a new understanding of how to create persuasive feedback.

Models for argumentation

The views represented in Toulmin's model for argumentation originally build on how courtroom arguments are structured, but are used broadly in the fields of rhetoric and communication today. The first three elements of the model: claim, grounds, and warrant are considered as the basic components of practical arguments, and describe the aim of the argument (the claim), the evidence (the grounds), and the underlying assumption, that the receiver must agree upon in order to accept the grounds (the warrant) (Toulmin, 1958). For a quick analysis of a fabricated usability problem, claim, grounds and warrants might look as shown in Figure 1.

A quick analysis of any argument will reveal if the underlying warrant is generally agreed upon, or if it needs further backing to make the argument stronger. In the fabricated example from Figure 1, an evaluator might add '15% of the users of this application are colour blind, which makes them an important group of customers' to back up the warrant, making the argument stronger and more persuasive.

The three modes of persuasion

In his classical work 'Ars Rhetorica', Aristotle describes the three modes of persuasion: ethos, logos and pathos (Aristotle, 1991) as the basic means of delivery of an argument, see also Figure 2.

Ethos is described as the trustworthiness of the personal character, and can be obtained through display of skills and wisdom, virtue or goodwill (Aristotle, 1991). In the usability community this notion might be translated as a confident and experienced evaluator who makes insightful remarks and conclusions about the system being evaluated. In the fabricated example above, mentioning having collected user demographics from the Marketing department, suggests that the evaluator is thorough and has domain knowledge, thus boosting his ethos.

Logos is the logical appeal (Aristotle, 1991), and is often based on the use of quantitative scientific or empirical data. Applied in the usability community this might mean using log data or statistical data from a user test as backing for a usability problem, as the '15%' in the example shows.

Pathos is explained as the moving of the receiver's emotions (Aristotle, 1991), and can be accomplished by for instance using words with strong positive or negative connotations, metaphors, stories, or a passionate style of delivery. Applied in usability evaluation this might mean showing videos of users struggling with an application or even letting the developers experience the problems themselves. In the example from Figure 1 the producer of the feedback could add a few words to describe the emotional impact on the colour blind users to influence the receiver: 'the distinction between red and green buttons is not perceived by people who are colour blind, and leaves them clueless and confused as to how to use the system'.

Next, we present and re-examine two field studies in order to test whether a feedback format's success or failure can be explained by a comparison with the two models of argumentation. We aim to find out whether feedback formats that present information in a manner that corresponds with Figures 1 and 2 are more successful in making developers acknowledge UPs than the formats, which do not.

Field studies

We examine two studies, which were conducted separately in two Danish companies. Each study

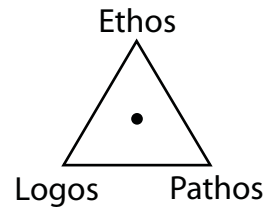


Figure 2: The figure shows how a well-balanced argument (the black dot) is balanced between the ethos, logos and pathos modes of persuasion.

was designed to evaluate the strengths and weaknesses of various types of usability feedback. The purpose of this re-examination is to study the combined findings of the field studies, not to directly compare the separate results of the two studies. A brief summary of each field study is presented below, more details can be found in (Nørgaard & Hornbæk, 2007; Høegh, 2007). An overview of the studied feedback formats is presented in Table 1.

Study A

Study A aimed at understanding how various presentation formats influence developers' assessment of usability claims. The study took place at a company that provides online services for job searching. Three developers, all familiar with receiving usability feedback as part of their jobs, participated in the study.

The study investigated five feedback formats, which represented different approaches to presenting and/or arguing for usability claims (see Table 1). The selection of formats was based on a review of work on feedback formats, such as (Coble, Karat, & Kahn, 1997; Dumas, 1989; Dumas, Molich, & Jeffries, 2004; Hornbæk & Frøkjær, 2005; Jeffries, 1993; Kennedy, 1989; Mills, 1987; Nayak, Mrazek, & Smith, 1995; Redish, Bias, Bailey, Molich, Dumas, & Spool, 2002; Schell, 1986) and an informal survey about preferred feedback methods on an online professional forum.

A think aloud test identified 40 usability problems. These were all presented to the developers using five feedback formats. Each developer individually answered five questions regarding the feedback's quality such as usefulness, persuasive power and clarity. Next, the developers worked with the problems for 12 weeks and then reassessed the quality of the feedback. Finally, the developers were individually interviewed about the five formats. The study found that developers, before having worked with the UPs, rated the formats which included videos and redesign proposals to be the most useful. After having worked with the UPs no significant differences between

Study	Format and content	Result
A	Multimedia. Linked html-documents with a problem description, a screen dump of where the problem occurred, video clip showing user interaction, a graphical and textual redesign proposal, a graphical illustration of severity.	Very convincing but time consuming
	Problem list. Problem description, severity rating.	Useful for simple problems otherwise too brief
	Redesign proposal. Problem description, redesign proposal, a justification of the redesign proposal.	Convincing
	Human-centred story. A narrative that describes a usability problem, the context and the user's responses and feelings.	Not convincing, difficult to understand
	Screen dump. Screen dumps annotated with information about where the problem occurred, a brief problem description and a textual redesign proposal	Convincing
B	Redesign workshop. Oral presentation of problem and severity rating, video clip showing user interaction, illustrated redesign proposals and face-to-face discussions with interaction designers.	Very convincing but very time consuming
	Report. Brief problem description, severity rating, reference to how many users experienced the problem, transcribed log files	Useful for simple problems otherwise too brief
	Self-experience. Developers follow a list of tasks designed to lead them through troublesome areas in the application	Very convincing but very time consuming

Table 1: The eight feedback formats, which were investigated in studies A and B. The table presents a description of the formats' content and a brief description of how it was appreciated by developers.

the ratings of formats were found.

Study B

The purpose of field study B was to evaluate how various formats of feedback from usability evaluation impact the developers' understanding of their software. The study was carried out at a company, which develops advanced diagnostic and data analysis products for telecommunication companies. The systems used in the field study were complex systems designed to handle large amounts of data.

During the study seven developers from two software teams, who were experienced with receiving usability feedback, were asked to write down what they considered to be the top five strengths and weaknesses in their software. The software was, respectively, an administrative system and a

presentational system. The software was evaluated using the think-aloud protocol and the results were analyzed, resulting in a total of 70 usability problems. These problems were presented to the two teams of developers using three different feedback formats (see Table 1). After one software development iteration the developers were asked to write down their current view on the top five strengths and weaknesses. The study found that developers highly appreciated the redesign workshop that offers the possibility to watch videos of user interaction, to discuss the findings with usability experts and to work together to solve the usability problems. Further, the self-experience format, which let developers experience the usability problems themselves, showed to be both persuasive and useful for providing an understanding for the problem.

Next, we discuss the results from the two field studies and compare the argument structure of the feedback formats with reference to the Toulmin and Aristotle models.

Findings

When understanding feedback as argumentation and not just as a presentation of results, we need to look to argumentation theory for explanations of what makes feedback persuasive.

Below we compare different feedback formats' argument structures to the concepts from Toulmin and Aristotle. We do so to investigate if concepts from argumentation theory can explain why some formats apparently present information in a way which is more likely to make developers acknowledge a UP than other formats. Our findings suggest that formats, which are regarded the most persuasive by developers, more closely follow the structure of the two argumentation models than formats which are considered less persuasive. Further, our findings suggest that the two formats which understand feedback as a learning process rather than a static deliverable, most effectively make developers acknowledge UPs.

Findings related to Toulmin's model for argumentation

Table 2 presents our analysis of how different argumentation styles manifested in the feedback formats are related to Toulmin's model for argumentation.

Since we look at feedback formats and not at individual descriptions of problems we have analysed how a format generally argues for a claim, and not how every single problem is specifically argued. For instance: all redesign proposals hold the warrant 'the proposal can actually be implemented in the system', but a specific proposal might also hold warrants such as 'people who are colour blind should be able to use the application easily'. Our analysis seeks to identify how a format argues for a problem in terms of claim, grounds and warrants. However, since not all formats present a concept equally thoroughly, the table distinguishes between concepts that are clearly presented (++), occasionally/vaguely presented (+), not presented (-) and meanings, which the receiver must deduce himself (?). For the purpose of analysis we have identified the claim as the claimed usability problem and the grounds as any evidence presented to support the claim. We identified underlying warrants by asking 'what basic view must the receiver agree upon to accept

the grounds and the claim?' A warrant for, say a redesign workshop, might thus be 'the video clip shows the truth', 'the expert is right' or 'personal observations and experiences are reliable'.

Below we present and discuss four primary results of the analysis: (1) how some feedback formats present one claim at a time, whereas others present several claims together; (2) how only some formats support all claims with grounds; (3) how some formats leave it entirely to the receiver of the feedback to deduce both claims and grounds; and (4) how none of the formats attempt to explain the underlying assumptions on which their argument builds.

Presenting more than one claim

Table 2 shows how the different formats present between one and three claims, for instance one claim about a UP and one about its severity. We expected that if a format presents multiple claims, and the developer rejects one of these, the rest of the claims might end up being rejected too, leaving the problem unacknowledged. This seems not to be the case however. The data shows no connection between how many claims a format presents and how well developers receive it. The interviews from study A state that receivers of redesign proposals do not always agree with the redesign proposal itself (the secondary claim), but this does not lead them to reject the relevance or nature of the usability problem (the primary claim). A developer elaborated on this phenomenon: 'Even though I do not necessarily agree with them, redesign proposals show that the evaluator has considered the reported problems thoroughly, and that makes the feedback more credible'.

Data suggests a hierarchy of claims, where the UP is the primary claim, and where redesign proposals, severity ratings and such serve as supportive claims that elaborate on, support, or facilitate a solution of the primary claim. Our findings suggest that the dismissal of a secondary claim does not necessarily lead to developers dismissing the nature or relevance of the usability problem. Consequently, supporting a usability argument with one or more secondary claims does not seem to endanger the persuasive power of usability feedback.

Some claims lack grounds

Some formats present several grounds for their claims where as others only present minimal grounds and perhaps only grounds for some of the claims. We expected that the most convincing feedback formats were the ones that presented a high number of grounds to support their claims. However, data suggests that there is no connection between a high number of grounds and de-

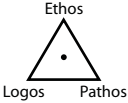
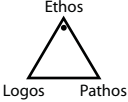
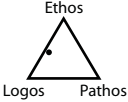
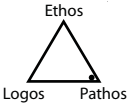
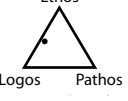
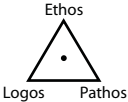
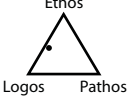
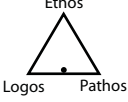
Format	Argumentation		Example
	Claims, grounds, and warrants	Modes of persuasion	
Multimedia	Claims: ++,+,+ Grounds: ++,+,+,- Warrants: -		A textual problem description presents a usability claim: <i>'The results do not show the user where the job is placed geographically'</i> . Grounds are provided by a screen dump, video clips and references to user behaviour. An unsupported claim is made about the severity of the problem. A textual and illustrated redesign proposal is presented as a claim.
Problem list	Claims: ++,++ Grounds: ++ Warrants: -		A textual problem description presents a usability claim: <i>'The geographical category 'other countries and Greenland' is not specific enough. Severity level 3: The site appears not to seriously target people who wish to work outside Denmark' and makes the site seem superficial.'</i>
Redesign proposal	Claims: ++,++ Grounds: ++,+,+ + Warrants: -		A textual problem description presents a usability claim: <i>'Several users overlook the results from the initial search and repeat the search once entering the results page'</i> . A textual and illustrated redesign proposal is presented as a secondary claim and a justification of the solution is presented as grounds.
Human-centred story	Claims: ? Grounds: ? Warrants: -		A one page scenario presents a fictitious use situation and focuses on depicting users' dialogue and emotions. [Excerpt from human-centred story:] <i>'I think I will type 'biotechnology' in this field...and it has to be situated in Copenhagen'. She types the word and chooses 'Copenhagen' from the drop down menu. Then she clicks the search button. 'Nothing? How can they have nothing at all?' she moves the cursor and chooses 'Entire country' as the geographical area instead. Then she re-clicks the search button. 'Try have a look...there is nothing here?' She looks questioningly at Maria.</i>
Screen dump	Claims: ++,++ Grounds: ++ Warrants: -		A textual problem description presents a usability claim: <i>'Several users overlooks or chooses not to use the possibility to search within categories or combine more categorical searches'</i> . A screen dump is presented as grounds for the claim.
Redesign workshop	Claims: ++,++ Grounds: ++, ++,?,?,?,? Warrants: -		A textual problem description presents a claim: <i>'The system performance is too slow. The users get frustrated while waiting, or sometimes misinterpret the situation. One user thinks his interaction with the system was not registered by the system'</i> . A redesign proposal presents a secondary claim, and video clips showing users experiencing the UP are provided as grounds. A discussion between participants facilitated the decision of whether or not to acknowledge the UP.
Report	Claims: ++,++ Grounds: ++,+,+ Warrants: -		A textual problem description presents a claim: <i>'Lack of feedback on system status. The user is not sure if the system is loading data or not'</i> . The claim is grounded in log files, a severity rating, and an illustration of how many users experienced the problem.
Self-experience	Claims: ? Grounds: ? Warrants: -		The developer is presented with the same tasks as the ones used to guide users during the think aloud test. No specific claims or grounds are presented regarding usability problems, the developer has to deduce these himself while working with the tasks.

Table 2: An analysis of how different feedback formats structure their argumentation. The table is based on concepts from Toulmin and Aristotle. The table distinguishes between concepts that are clearly presented (++) in the feedback, occasionally/vaguely presented (+), not presented (-) and concepts, which the receiver must deduce himself (?). For example, the multimedia format presents three clear claims. To support these, the format presents two clear grounds, one vague ground, and one ground, which the developer must deduce for himself. One claim has no grounds. Further, the format does not make any warrants explicit.

velopers acknowledging the usability claim.

For example both redesign workshops and multimedia presentations present multiple grounds for their claims using video clips, screen dumps, expert's opinions, personal observations etc. to deliver evidence for a claim. These formats are highly valued by developers and suggest a connection between presenting a high number of grounds and getting the claim acknowledged. However, two other formats, redesign proposals and self-experience, only present minimal grounds for their claims, solely relying on experience or a justification for a redesign proposal to prove their usability claims. Since both redesign proposals and self-experience formats are also highly valued by developers, the scarcity of the evidence presented to support their claims does not seem to effect whether the problems are acknowledged.

Based on these findings we might conclude that it is not possible to predict whether a problem will be acknowledged by looking at the number of grounds a format presents to support the claim. However, study A showed some interesting similarities between low rated usability problems. Generally, and regardless of format, all developers gave a low rating to problems, which they were not convinced about, had difficulty understanding or which they needed more contextual information about. Observations from study B suggest the same pattern. As an example, one developer rated a problem regarding a malfunctioning search field very low, and explained that he needed to know what exactly was typed in the field before he would acknowledge the problem. It might not be a problem with the search field, he explained, but could simply be the user typing in words that had no match. The similarities between the low rated UPs, as problems that are hard to understand or unconvincing, might be an indication that the amount of grounds presented or the convincing power of the grounds does matter. We hypothesize that different grounds weigh differently, and that one 'heavy-weight' ground may easily be as convincing as two or more 'low-weight' grounds. Further, based on observations from study B, which suggests that developers do not always agree on what is a heavy weighing ground, or whether or not to acknowledge a UP, we hypothesize that different grounds also weight differently on different people. The analysis suggests a pattern where self-experienced grounds weigh heavier than explained ones, meaning for example that watching a video of users interacting with the system weighs heavier than an expert's opinion. The ratings of the two learning-oriented formats: the redesign workshop and the self-experience format, suggest that experiencing

or discussing UPs efficiently help developers acknowledge UPs.

To further investigate our hypothesis that not only the number but also the weight of grounds matters, we suggest looking deeper into what constitutes a heavy weighing ground in order to help producers of written feedback support their claims most persuasively.

No explicit claims or grounds

Some formats are very explicit in their description of claims and grounds, whereas others only present vague descriptions or leave it to the receiver of the feedback to deduce the needed information. For example, a problem description from the problem list format is a very explicit claim, whereas the self-experience format leaves it to the developer to deduce both claim and grounds from his own observations. We expected that formats that rely on developers to observe, analyse and interpret the feedback in order to deduce claims or grounds on his own, might be more persuasive, since 'seeing things for yourself' and 'figuring things out for yourself' is more persuasive than being told.

The data suggests that formats that rely on developers to deduce claims or grounds can be, but are not necessarily, successful. The formats self-experience and human-centred stories both build on the idea that the receiver of the feedback must analyse data and deduce the claim or the grounds on his own. But while the former is highly appreciated the latter is criticized by developers. Users of human-centred stories describe the format as highly confusing and seem unsure about how to use them. One developer states that it is surprising how one can read a one page description of a problematic use situation, and still know very little about the problem. This suggests that the deduction of claims and grounds from text is difficult. In contrast, developers, who worked with the self-experience format, expressed that they found it easy to deduce claims and grounds, and found the format convincing since they had gained an understanding for the user's perspective.

Redesign workshops and the multimedia format also rely heavily on developers' ability to deduce claims and grounds. They present video highlights, presupposing that developers are able to deduce usability claims and grounds from a video. Observations from both studies suggest that developers are able to deduce claims and grounds from videos, and confirm that developers generally appreciate the possibility to interpret and conclude claims and grounds for themselves rather than having somebody do it for them.

The data suggests that developers are capable

of and feel confident analysing data to deduce claims and grounds from videos as well as their experiences from a self-experience session. And that feedback formats using this type of argumentation are considered very persuasive. The exception is the task to deduce claims and grounds from a human-centred story. This proves very difficult, possibly because in order to deduce claims and grounds, the developer needs to draw from multiple pieces of contextual information, which the human-centred story seems not to provide.

Providing feedback using video highlights (as in the multimedia or redesign workshop formats) or using the self-experience format is however a costly affair, since they are time consuming to produce, plan and use. They are regarded as very persuasive by developers though, and might be reserved for feedback on very controversial problems. For claims, which are not highly controversial, we expect that an explicit and well-structured written argument is sufficient to make developers acknowledge the relevance and nature of a UP. The learning-oriented formats: the redesign workshop and the self-experience format stand out as being highly successful in making developers deduce claims and grounds on their own. We hesitate to explain this success simply with argumentation theory. Since developers seemingly acknowledge some UPs more easily when they have worked with them or discussed them with others, we suggest that the persuasive power of learning-oriented formats should be studied further from a learning point of view.

No explained warrants

The fourth finding has to do with warrants. Warrants are, as we have previously stated, underlying and implicit assumptions that must be agreed upon in order for the argument to have an effect. We have analysed the feedback formats to identify whether their argument structure resembles Toulmin's model for argumentation. To construct a convincing argument, an evaluator would analyse claim and grounds to identify which warrant they rest on. If the warrant is controversial, or can easily be challenged, the evaluator would need to present specific grounds to support it. The warrant will then become a claim on its own, a claim that has its own grounds and warrants. An argument analysis will form a ladder of continuous arguments. Eventually the analysis will reach a warrant that is so basic, that it does not need to be supported. To give an example, the warrant 'the company depends on paying customers' would most likely be considered unchallengeable for a profit-oriented company.

In our analysis of how the different feedback formats argue for a claim, we looked for such ladders

of claims and grounds. However, no format presents a ladder of arguments. This suggests that no attempts have been made to identify which implicit assumptions lie behind a claim, nor to present any support for such warrants in order to improve the chances of the UPs being acknowledged. As a couple of examples, let us consider what warrant lies behind using edited video clips as grounds for a claim, such as it is done by the multimedia and redesign workshop formats. In study B, a claim about the system responding too slowly, was presented. This claim was supported with video clips showing users waiting for system response. To accept the claim and grounds a developer would need to recognize that what was shown by the video clips was true and representative for the user experience. In another example, redesign proposals are used by the formats redesign proposals, multimedia presentation, and screen dumps (the redesign workshop format does not present explicit redesign proposals but provides the opportunity to discuss redesign ideas. With respect to argumentation this is different from the other formats). A proposal for a redesign claims that the implementation of a certain idea will solve the usability problem. To support this claim some formats present grounds in the form of illustrations and justifications. However, to accept the argument one must also agree that the redesign proposal can actually be implemented in the system, and that it will indeed solve the problem. In these examples, developers might not agree with the video clips being representative or a redesign proposal being implementable, and might reject the entire claim.

We expected to find that warrants, which could easily be challenged, such as 'the expert is right' and 'the videos are representative', to result in a rejection of the claim. This seems however not always to be the case. Developers did not question the video clips, but considered them to be the next best thing to actually attending the test. Developers who attended the redesign workshop did not question the expertise of the usability expert either. However, when we looked beyond the feedback formats and into specific problems and specific warrants, we discovered that developers do not always agree with warrants. Sometimes they do reject claims because of unsupported warrants. For example one developer rejected the claim that the use of inconsistent terminology was a problem. He recognised that a service, offered by the system, was called by three different names, but he completely rejected the warrant that users will be confused about inconsistent terminology. In this case the evaluator might have explained how users expected that the three different words actually described three different services. And how they, as a consequence of the

inconsistent terminology, became confused about the exact differences between the three seemingly related services, and lost confidence that they were capable of understanding and using the application properly.

In another example, a developer refused to acknowledge a UP because he felt that a task was explained in a way that had caused the user's confusion. Thus, he was uncertain if the problem was actually caused by the system, or if it was caused by how the task was explained to the user. In this case, the developer refused the warrant that the test was conducted in a trustworthy manner, and dismissed the entire UP.

As we have seen, data from both studies suggests that sometimes it does matter whether the receivers of the feedback agree on a claim's warrant. Some UPs are in fact dismissed because developers do not accept the argument's warrant. Accordingly, we urge practitioners to analyse their arguments in order to understand the warrants. We argue that the producer of feedback could improve the persuasiveness of any claim by identifying which warrants the claim builds on and present specific grounds for any warrants which can easily be challenged.

We have identified how presenting both primary and secondary claims do not seem to weaken an argument. How some grounds seemingly weigh heavier than others. How relying on developers to deduce claims and grounds on their own, seems a very persuasive way for feedback to present an argument. Finally, we have identified how unsupported warrants may cause developers to dismiss a claim.

Next, we discuss important findings related to the three modes of persuasion.

Findings related to the three modes of persuasion

In Table 2 the triangular model depicts how each feedback format balance between the use of ethos, logos and pathos. Below we present and discuss the following three findings: two formats rely solely on one mode of persuasion: four formats rely on two modes of persuasion, and two formats show optimal balance between the use of ethos, pathos and logos persuasion.

Relying entirely on one mode of persuasion

Our analyses show that the problem list and the human-centred story formats both rely on one single mode of persuasion.

Since the problem list format rarely presents any grounds for its claims, its power of persuasion

is entirely based on the producer's ethos. If the receiver of the feedback for some reason would reject the producer's ethos (for instance his role as an expert) the claim would be easily dismissed. Accordingly, we hypothesized that any format, which relies on one single mode of persuasion, would be less persuasive than formats, which rely on several modes of persuasion. However, the problem list format seemingly works well for presenting uncontroversial UPs, which confirms that uncontroversial UPs do not need to be fed back with meticulous attention to how the argumentation is structured and delivered. The interviews in study A showed how one developer specifically used severity ratings as a means to estimate the producer's ethos. If the developer agreed with the producer's severity rating, he would gain respect for the producer's expertise, and forthcoming claims were more likely to be acknowledged. The finding suggests that receivers of feedback seek for ways to estimate or test the producer's ethos. Since developers rated the problem list format as sufficient for a whole range of problems, we acknowledge that the producer's ethos in many cases is in fact sufficient to make developers acknowledge a UP. However, for more controversial problems, for example problems that would require a lot of work on behalf of the developer, relying solely on ethos might not prove satisfactory. Several problems, which were fed back using the problem list format, were rejected because the developers wanted more information about a problem before being willing to acknowledge it. This confirms that for some problems using ethos as the single persuasive mode is not sufficient.

Human-centred stories also rely on one single mode of persuasion. In their aim to influence developers through an emotional and dramatic narrative they solely rely on pathos to persuade the reader. Human-centred stories do not succeed, however. In one case the format backfired and resulted in one developer angrily refusing the usability claims; a problem with relying on pathos already described in the literature (Aristotle, 1991). Human-centred stories are an example of how difficult it is to use/create a feeling of pathos. Since pathos relies on influencing or manipulating the receivers' emotions, it can easily backfire. During an oral presentation the speaker can use pathos according to how the audience reacts, but a producer of written feedback has no idea how his reader may react to the text, and the use of pathos may end up having a different outcome than expected. With the heavy criticism of the human-centred story format in mind, relying solely on pathos in written feedback seems ill advised.

Relying on two modes of persuasion

Five of the eight feedback formats rely on two

modes of persuasion. Redesign proposals, screen dumps and the report formats all rely on ethos and logos arguments. These feedback formats support some of their claims with logos arguments and some they leave unsupported, hence relying on the producer's ethos. Our data show that developers value these formats highly. As an example all developers in study A characterize redesign proposals as inspiring and explain that since the producer has taken the time to develop and illustrate one or more redesign proposals the feedback has higher quality and seems more convincing. The effort put into producing the feedback seemingly reflects positively on the producer's ethos. Data show that developers respond positively to the logos mode of persuasion, and we hypothesize that this further boosts the producer's ethos so not all claims need to be supported specifically, but are acknowledged because of the producer's experience.

Self-experience relies on logos and pathos modes of persuasion, and since the producer is not very visible in the feedback, the matter of ethos seems less important for the self-experience format. Data show that the self-experience format is highly appreciated, and developers explain that it helps them get a clear understanding of the use situation and problems with the interaction. Again, the self-experience format is closer to a learning process than a written deliverable, and the pedagogical nature of the format may thus also explain its persuasive power.

Based on our data we conclude that formats, which rely on two modes of persuasion are considered more persuasive than formats, which rely on a single mode of persuasion. For example, the report format is basically the problem list format equipped with log files and statistics, which fall under the logos mode of persuasion. This shift from solely using one mode of persuasion (problem list) to using two modes of persuasion (report) appears effective since the report seemingly gets less problems rejected than does the problem list.

It is quite clear that out of the four formats which use two modes of persuasion, the self-experience format is the most costly to plan and produce. It is furthermore very time consuming to use, and developers complained that for an average sample of problems, which include problems that are both easy and difficult to understand, and problems that are both controversial and uncontroversial, using the self-experience format is simply too time consuming, since it requires the developer to work with and experience every single UP. Despite it being a format that effectively makes developers realize the relevance and nature of

UPs, the self-experience format does have major drawbacks in terms of costs both on behalf of the producers and on behalf of the developers, and should probably be reserved for controversial UPs.

The well-balanced argument

Two formats seem to balance admirably between ethos, logos and pathos. Multimedia presentations and redesign workshops use expert's opinions and problem descriptions which mainly rest on ethos. They use screen dumps, justifications, severity ratings and statistics, which mainly rests on logos. And finally they both use video highlights most successfully to add pathos to their argumentation. In this respect they present a perfectly balanced argument. However, some developers described the use of videos as a paradox. On the one hand they appreciated the rich contextual information that videos present, but on the other hand they were annoyed by the tediousness and slow nature of the medium. Despite the videos only being between 30 seconds and 6 minutes long, several developers explained that this was too long. Two developers suggested adding a fast forward button on the video, enabling browsing the video and the ability to fast forward to the 'point of the problem'. Quite paradoxically compared to their expressed need for contextual information, they found it too tedious to watch, say, 30 seconds leading up to the actual usability problem.

Despite multimedia presentations and redesign workshops as being rated the most persuasive and appreciated formats, they are fairly costly to plan and produce. Even the time one needs to set aside to watch the videos is considered too much by the developers. In conclusion, we hesitate to simply recommend these two formats as the best. We suggest practitioners cooperate with domain experts to estimate the controversial nature of a problem. Based on this estimation, decisions could be taken about whether a UP is best presented using formats that only rely on two modes of persuasion, or if the controversial nature of the problem calls for a feedback format, which uses all three modes of persuasion.

Before we reach the conclusion, let us just briefly return to the paradox of wanting a lot of detailed information and still being able to use the feedback very quickly. Developers from study B emphasized that the amount of information presented by a feedback format had a great influence on whether the feedback got accepted or not, regardless of the argumentation or the format. They agreed that they would often ignore feedback, which presented many problems in great detail simply because the amount of information was overpowering. Since it would take vast resources

to present every identified UP with video clips, or to discuss even the smallest usability problem in a redesign workshop, some feedback formats have a natural upper limit for how much information they can present. However, some formats, such as the report, do risk overpowering developers with information, because it is fairly easy to list a large amount of problems or to include a huge amount of detailed log information.

We recognize that developers experience the paradox between wanting a detailed overview of the UPs and being able to get that information very fast. Accordingly, we suggest that working specifically with the structure of arguments and modes of persuasion will help address the need for detail and clarity while keeping the feedback to a manageable size. We recommend that feedback is presented in a way that allows the reader to examine details (such as videos or log files) if needed, but also facilitates browsing or a quickly-read overview of the problems. We acknowledge that for some UPs the pedagogical qualities of discussing and experiencing problems might be crucial for whether a developer will acknowledge the problem or not.

To sum up, our data suggests that the more modes of persuasion a format uses, the more persuasive it is. However, the formats that best cover the three modes of persuasion are also the formats that require a lot of resources from both producers and developers. Developers specifically state that there is an upper limit for how many problems they can work with using time-consuming formats. Accordingly, we advise producers of feedback to assess if a problem is so controversial that it needs to be fed back using the multimedia or the redesign workshop formats. Perhaps it could equally successfully be presented using formats, such as redesign proposals, screen dumps and the report format, which only use two modes of persuasion.

Conclusion

Studying usability is widely accepted as an integral part of developing software. However, how to deliver the results from usability evaluations is an area mostly left unstudied. We argue that understanding feedback from usability evaluations as an argument for a usability problem might prove more constructive than simply considering feedback as a presentation or description of a usability problem, as most literature on the subject does, see for instance (American National Standards Institute, 2001; Dumas & Redish, 1993; Mills, 1987; Redish, Bias, Bailey, Molich, & Spool, 2002; Rubin,

1994).

As shown in Table 2 none of the eight formats studied in this paper try to make warrants explicit. In retrospect, it might be questioned whether our claim about usability feedback being an argument is correct, when such a crucial part of the argument is seemingly missing. However, despite the fact that the formats we have studied do not seem to elaborate on warrants, there are still implicit warrants behind the presented usability claims. And despite both claims and grounds often are being clearly presented, we have encountered several UPs that were dismissed because the receiver did not accept the warrant. With this in mind, we argue that understanding feedback from usability evaluations as arguments, and focussing on creating written feedback as well-structured arguments will help to improve the persuasiveness of the issues being presented.

On the subject of argumentation structure, we recommend that producers make sure that claims and grounds are clearly described. Also, the warrants behind the argument need to be identified and supported if they can easily be challenged. The study suggests that the persuasiveness of a format relates to how many modes of persuasion it uses. Constructing arguments based on ethos and logos should be sufficient for most problems, but for controversial UPs, evaluators are advised to look to the use of videos, or to use learning-oriented processes to engage developers emotionally.

In order to choose the most cost-effective feedback format, we advise that feedback producers estimate the controversy of UPs together with someone who has specific domain knowledge. To focus resources where they are most needed, uncontroversial problems could be fed back in a relatively simple report style which will allow more time for producing thorough feedback on controversial UPs, using for instance multimedia presentation or learning-oriented processes, such as redesign workshops.

We acknowledge that viewing written feedback isolated from the context in which it is delivered and used, might not be an accurate picture of how written feedback is used. In industrial settings written feedback is often combined with oral elaborations, and possibilities to ask questions about and discuss the findings. Accordingly, considering the context in which written feedback is delivered and used, is important for many aspects of the quality of feedback. For complicated and controversial usability problems the possibility to discuss or experience UPs is seemingly an important contributor to understanding and acknowledging a problem. However, constraints on

time and money in software development do not lend opportunity to plan and execute feedback as cooperative learning processes for each and every usability problem. Thus, the value of short and clear written feedback, as a means to provide information about usability problems, is unquestionable.

Some of our findings seem, however, related to pedagogical aspects of feedback rather than aspects strictly related to argumentation. For example, developers seemingly feel confident and convinced about deducing claims or grounds from feedback on their own, a practise that seemingly understands feedback as a learning process. While understanding written feedback as argumentation may explain why some problem descriptions are more convincing than others, we also suggest studying feedback in terms of pedagogy such as viewing feedback as cooperative problem solving or as a learning process. For future work we expect to look into how theories of learning can improve the entire process of feeding back results from usability evaluations and facilitate the solution of crucial usability problems in software.

This paper is based on earlier work, which was not specifically designed to test how successful feedback formats map to argumentation theory. Further, the two studies are based on data from only 10 developers. Accordingly, we suggest that future work study how evaluators can produce feedback as persuasive arguments and how developers receive and use such feedback. Since the grounds presented to support a claim are apparently assessed differently by different developers, we also suggest specifically looking into what criteria developers employ to assess the quality of an argument, and the weight of the grounds.

To sum up, we claim that developers need to acknowledge the relevance of usability feedback and the existence of a UP before they will act on it. Other work on how to feed back results from usability evaluations provide recommendations such as presenting usability problems in a respectful tone of voice, avoiding technical jargon, presenting positive findings and so forth, see for instance (Dumas & Redish, 1993). We recommend understanding written feedback from usability evaluations as an argument for a series of usability problems. Accordingly, thinking in terms of claims, grounds, warrants and modes of persuasion will help evaluators produce persuasive arguments for the existence and relevance of usability problems.

Acknowledgements

We thank our colleagues Erik Frøkjær, Kasper Hornbæk, Effie Law and Lene Nielsen for their insights and valuable input. The work is part of the USE-project (Usability Evaluation & Software Design) founded by the Danish Research Agency through the NABIIT Programme Committee (Grant no. 2106-04-0022).

References

- American National Standards Institute The Common Industry Format (ANSI/NCTS-354-2001), New York, (2001).
- Aristotle; 'On Rhetoric': A Theory of Civic Discourse, Kennedy, George A. (trans./ed.), Oxford University Press, New York/Oxford, (1991).
- Boivie, I., Åborg, C., Persson, J., & Löfberg, M.; Why Usability Gets Lost or Usability in in-House Software Development, *Interacting with Computers*, 15, 4 (2003).
- Coble, J. M., Karat, J., & Kahn, M. G.; Maintaining a Focus on User Requirements Throughout the Development of Clinical Workstation Software, *Proceedings of the ACM Conference on Human Factors in Computing*, 22-27.March 1997, (1997).
- Dumas, J.; Stimulating Change Through Usability Testing, *SIGCHI Bulletin*, July 1989, 21, 1 (1989).
- Dumas, J., Molich, R., & Jeffries, R.; Business: Describing Usability Problems: Are We Sending the Right Message?, *Interactions*, 11, 4 (2004), 24-29.
- Dumas, J. & Redish, J. A.; *Practical Guide to Usability Testing*, Ablex, Norwood, (1993).
- Høegh, R. T.; The Focus of Current HCI Research in Usability, *Proceedings of the 7th Asia-Pacific Conference on Computer-Human Interaction, APCHI 2007, Taipei, Taiwan, (2006)*.
- Høegh, R. T.; Software Development and Feedback From Usability Evaluations, To appear in *Proceedings of ITAIS 2007, Venice, Italy, (2007)*.
- Hornbæk, K. & Frøkjær, E.; Comparing Usability Problems and Redesign Proposals As Input to Practical Systems Development, *ACM Conference on Human Factors in Computing Systems*, (2005).
- Jeffries R.; Usability Problem Reports: Helping Evaluators Communicate Effectively With Devel-

- opers, in *Usability Inspection Methods*, (1993), 273-294.
- John, B. E. & Marks, S. J.; *Tracking the Effectiveness of Usability Evaluation Methods*, *Behaviour & Information Technology*, 16, 4/5 (1997), 188-202.
- Kennedy, S.; *Using Video in the BNR Usability Lab*, *SIGCHI Bulletin*, 21, 2 (1989), 92-95.
- Law, E.; *Evaluating the Downstream Utility of User Tests and Examining the Developer Effect: A Case Study*, *International Journal of Human-Computer Interaction*, 21, 2 (2006), 147-172.
- Mills, C. B.; *Usability Testing in the Real World*, *SIGCHI Bulletin*, 18 (1987), 67-70.
- Molich, R.; *User-Friendly Computer Systems* (in Danish), Teknisk Forlag, (2000).
- Nayak, N. P., Mrazek, D., & Smith, D. R.; *Analyzing and Communicating Usability Data*, *SIGCHI Bulletin*, 27, 1 (1995), 22-30.
- Nørgaard, M. & Hornbæk, K.; *What Makes a Developer's Heart Tick? Characterizing Effective Feedback From Usability Evaluation*, Technical report from Copenhagen University Dept. of Computer Science, <http://www.diku.dk/publikationer/tekniske.rapporter/rapporter/07-01.pdf> (2007).
- Redish, J., Bias, R. G., Bailey, R., Molich, R., Dumas, J., & Spool, J. M.; *Usability in Practice: Formative Usability Evaluations - Evolution and Revolution*, *Proceedings of the CHI*, April 20-25, Minneapolis, Minnesota (2002).
- Rubin, J.; *Handbook of Usability Testing: How to Plan, Design and Conduct Effective Tests*, John Wiley & Sons inc., New York, 1994.
- Schell, D.; *Usability Testing of Screen Design: Beyond Standards, Principles, and Guidelines*, *Proceedings of the Human Factors Society 30th Meeting*, Santa Monica, CA, (1986), 1212-1215.
- Toulmin, S. E.; *The Uses of Argument*, Cambridge University Press, Cambridge, UK, 1958.

Can Eye Tracking Boost Usability Evaluation of Computer Games?⁷

Sune Alstrup Johansen

IT University of Copenhagen
Rued Langgaards Vej 7
DK-2300 Copenhagen S
sune@itu.dk

Mie Nørgaard

Copenhagen University
Universitetsparken 1
DK-2100 Copenhagen
mien@diku.dk

Janus Rau Sørensen

IO Interactive
Kalvebod Brygge 35-37
DK-1560 Copenhagen
januss@ioi.dk

Abstract

Good computer games need to be challenging while at the same time being easy to use. Accordingly, besides struggling with well known challenges for usability work, such as persuasiveness, the computer game industry also faces system-specific challenges, such as identifying methods that can provide data on players' attention during a game. This position paper discusses how eye tracking may address three core challenges faced by computer game producer IO Interactive in their on-going work to ensure games that are fun, usable, and challenging. These challenges are: (1) Persuading game designers about the relevance of usability results, (2) involving game designers in usability work, and (3) identifying methods that provide new data about user behaviour and experience.

Introduction

Broadly speaking, a great computer game is accessible and intuitive to use while being fun and challenging at the same time. In this respect it differs from office systems, which primary goals are fast, easy and efficient interaction. For both types

⁷ This paper was originally published for Workshop on Evaluating User Experiences in Games, 4.th April 2008, CHI2008, Florence, Italy.

of systems the element of usability is crucial. However, since the goals, the use, and the context of the two types of systems are different, usability evaluation methods used to test office systems often fail when applied to computer games. Thinking aloud while playing a first person shooter game for instance, proves practically impossible to many players (Nørgaard & Rau, 2007). Also, specific concepts such as game play and re-playability relate to games and not to office systems.

In this paper we discuss the use of eye tracking as a means to address some of the usability related challenges faced by the computer games producer IO Interactive. Since our work is in progress we cannot report any results as to the successes or failures of using eye tracking to support and facilitate the usability work in the development of computer games. We will, however, argue why we expect that the use of eye tracking will address crucial challenges for usability work in this particular company, and perhaps other companies alike.

We do not claim that all IO Interactive's usability related challenges can be fixed by using eye tracking. Actually—and unfortunately—far from it. But we do expect eye tracking to tackle some of the challenges related to persuading game designers about the relevance of usability results as well as prove helpful when involving game designers in the usability work. Also, we expect eye tracking to improve the outcomes of retrospective think-aloud evaluations.

Related work

Much important work has been done on the development and use of usability evaluation methods that aim at evaluating office or web based systems. Such work include for example methodological studies and studies of how results can be described and reported, see for example (Hertzum, 2006; Hornbæk & Frøkjær, 2005; Jeffries, Miller, Wharton, & Uyeda, 1991; John & Marks, 1997; Redish, Bias, Bailey, Molich, Dumas, & Spool, 2002; Sears, 1997). In contrast, work on how usability in computer games is evaluated is limited. Helms Jørgensen explains the lack of descriptions of evaluation praxis with 'Microsoft [being] the only example of major game developers having seriously taken up usability approaches' (Jørgensen, 2004). Today, as the industry grows explosively and the competition increases, this is unlikely to be true. Several steps have been taken to facilitate the evaluation of usability in games during the years, see for example (Desurvire & Toth, 2004; Fabricatore & Rosas, 2002; Malone, 1982; Med-

lock, Wixon, Terrano, Romero, & Fulton, 2002). The indisputable value of this and related work aside, usability evaluation is still not necessarily a well-integrated part of the development of computer games, and usability practitioners are still in want for better methods and procedures to help them work specifically with the improvement of usability in games.

Next, we take a look at the specific challenges for the games producer IO Interactive, and discuss how the use of eye tracking might improve the impact that usability work has on computer games.

A company and its challenges

IO Interactive (IOI) is a Danish producer of computer games, and has since 1998 produced games such as the Hitman series. IOI, which is owned by SCi/ Eidos group, develops, designs and produces interactive entertainment for the major platforms on the global market. Though IOI's games are recognized for their game play, they are also known as being difficult to access for novices (Nørgaard & Rau, *User Testing in the Combat Zone*, 2007). And since the lack of usability is likely to have kept many users from getting value for their money or even purchasing the product in the first place, IOI has recently increased its attention on evaluating usability. Still, as has many before him, IOI's QA manager Janus Rau experiences that one thing is deciding to conduct usability work, quite another thing is making sure that the work has real impact on the design process.

IOI's usability challenges are described in a previous paper (Nørgaard & Rau, 2007) which placed them in five categories; justifying the costs, work procedures, user involvement, collaboration and alliances, and responsibility. The challenges interesting for this paper concern how to persuade game designers about the relevance of results from usability evaluations, how to better involve game designers in the usability work, and how to improve on some of the limitations of retrospective think-aloud testing.

The persuasiveness of results

To elaborate on these three challenges Rau explains that lack of understanding and knowledge about how usability work is conducted (such as the prejudice that usability professionals simply ask players what they like) could be a reason why many game designers are hard to convince about the relevance of usability results. Based on his work experience with IOI's game designers, Rau hypothesizes that presenting quantitative results using statistics, maps and graphs to supplement

more qualitative observation-based results might prove more persuasive than the qualitative results alone.

The involvement of game designers

Involving game designers in the usability work is one way of securing that important designer knowledge is fed into the evaluation work. Moreover, when game designers have been involved in usability work, they have a stake in it, and are more likely to acknowledge the results, Rau explains. But getting the game designers involved in usability work proves difficult. This may be because they do not feel they profit by the time spent watching videos of players and discussing usability issues.

The evaluation methods used

Finally, since the use of games is fundamentally different from that of office systems, not all traditional usability evaluation methods apply equally successful to computer games. The combination of observation, interview and questionnaire used in IOI might support a valuable dialogue with the player about the overall experience and specific incidents. However, it may not cover all games-related topics such as level of challenge, immersion and attention.

In the following we describe the method of eye tracking and discuss how the use of eye tracking may result in more persuasive usability results, how it may involve game designers in usability work, and how it may produce valuable results that other methods overlook.

Eye tracking

Visual perception is an essential part of users' interaction with games interfaces, and modern eye tracking equipment makes it possible to record and analyze parts of this process such as: Which elements are actually seen? And did modifications of the graphic design lead to the intended change in user gaze patterns?

Eye tracking has been criticized for being costly and tedious (Aaltonen, 1999; Schnipke & Todd, 2002). Difficulties calibrating the equipment to users with glasses, contact lenses, or even dark/brown eyes were common. Precision was low, and tiny head movements could jeopardize the validity of the recorded eye tracking data. State-of-the-art eye tracking equipment has solved most of these problems, and accurate recordings of eye movements can be made without obtrusive head-mounted cameras, or unnatural fixations of the

head in a stand. This is part of the reason, why the application of eye tracking technology in usability studies is clearly blossoming (Jacob & Karn, 2003; Poole & Ball, 2006).

Further, eye tracking has proved to be a valid method for discovering usability problems (Ball, Eger, Stevens, & Dodd, 2006; Goldberg & Kotval, 1999; Guan, Lee, Cuddihy, & Ramey, 2006), and is thought to provide an indication of the amount of cognitive processing required to interact with an interface (Rayner, 1998). Surprisingly, eye tracking as an evaluation method has not yet taken off into the area of computer games, as only few studies have been published, see for example (El-Nasr & Yan, 2006; Lin & Imamiya, 2006).

Getting new data from use of computer games

One particular qualitative method that has received much attention in relation to eye tracking research is the retrospective think-aloud method. Using the retrospective think-aloud method with eye tracking, usability researchers can let users interact with an interface without disturbing the interaction. This will make users focus more on the task at hand, and provide a more valid test situation. When a task is completed a video sequence can be shown to the user with an overlay of their eye movements. Recordings of eye movements have proven helpful to support the user in verbalizing his or her experiences and thoughts retrospectively (Hansen, 1990).

The experience of evaluation using retrospective think-aloud method is perceived as being subjectively more pleasant by the users (Ball, Eger, Stevens, & Dodd, 2006). Also, an increase in speed and focus on the task at hand has been observed, resulting in significantly higher task-completion rates than when using the conventional think-aloud method. This means that users are not distracted by the cognitive load they experience during traditional think-aloud tests. Traditional think-aloud testing makes it more difficult for them to concentrate on finishing their task. This is the reason why traditional think aloud testing often gets avatars killed (Nørgaard & Rau, 2007).

Apart from supporting retrospective interviews of user experience, eye tracking may also provide new data about user behaviour and experience that could add to the data gathered through current methods used by IOI.

During the development of a recent new game IOI experienced how players had difficulties getting by a team of snipers on a bridge. The questions concerned appearance of snipers: were the snipers spotted, but too big a challenge, or did

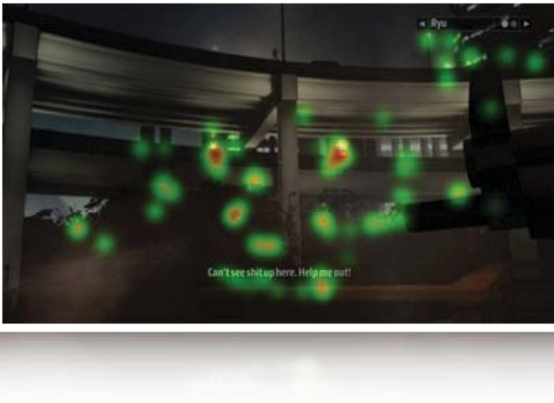


Figure 1: Heat map where players don't notice sniper.

the players not notice the snipers until too late? This question, and other questions related to attention, could be answered by measuring players' eye movements. Figure 1 shows a heat map for visual attention of players, where it is clear that the players don't notice the sniper in the middle top of the bridge.

Coloured areas indicate where players directed their attention (red is high attention, yellow is medium, and green is low attention). This problem could be solved by adding more light to the scene, like illustrated in Figure 2. The figure shows how players in a more well-lit version of the sniper attack were more likely to notice the sniper quickly.

Involving game developers with live eye gaze videos

During a test, the test moderator and observers are able to follow the player's eye movements on a screen. This possibility provides game designers and others with the possibility to get a better understanding of how players play and what is going on in the mind of a player during the game. We hypothesize that the access to such new information (and the engaging nature of a video with eye gaze patterns) might make game designers keener to get involved in usability work.

Gaze plots and heat maps increase persuasiveness

In addition to the qualitative data gathered by the retrospective think-aloud method, game evaluators can apply several quantitative measures to the test setup, since each session is completed without disturbances in the cognitive processes. This makes it possible to collect eye tracking data, and report on a wide variety of eye tracking metrics, e.g.



Figure 2: Heat map where sniper is quickly noticed.

- fixation duration that can tell if the user have difficulty in extracting information or finds an object especially engaging (Jacob & Karn, 2003).
- time-to-first-fixation that can reveal if an object or area has good attention-getting properties (Poole & Ball, 2006).
- fixation spatial density that can reveal inefficient search (Poole & Ball, 2006).

Such measures document, how the game interface is performing. The evaluator can generate illustrative visual output from the eye tracking data, such as heat maps (like Figure 1 and 2), and gaze plots that shows in which order players are directing their attention on screen. Such output are based on objective and quantitative measures and make it easier for evaluators to illustrate problems, document specific findings, or to convince the project team that results from usability testing are valid. Also, such outputs are generally considered persuasive for stakeholders such as developers, designers, and managers (Spool, 2006).

We suggest that eye tracking could be favourable to use for game evaluation, when focus and attention is vital for the game play, and when cognitive distractions can be devastating for 'player survival' in the game. Further, videos with overlay of eye movements support verbalization of experiences and thoughts in retrospective interviews. Eye tracking seem also promising as a means to involve game designers better in usability work since it offers completely new data about the use of a game. Finally, the quantifiable quality of the results of eye tracking suggest that game designers would consider results more persuasive that results derived using more qualitative methods.

Conclusion

In this paper we argue that the use of eye tracking to collect data during evaluations thanks to its quantitative nature might produce results that game designers at IOI consider more persuasive than for example results from traditional retrospective think-aloud tests. Also, videos showing eye gaze patterns or heat maps, that reveal what players see (and miss) on the screen, might involve game designers in usability work since they get the chance to get valuable information about for example the players' attention. Likewise, the use of eye tracking may improve the quality of retrospective evaluation since videos with eye gaze patterns may diminish a player's tendency to subsequently rationalize the gaming experience and the attention on the screen. Eye tracking may also provide new data about user behaviour and experience that current methods used by IOI miss, such as quantitative data about attention and orientation when playing a computer game.

References

- Aaltonen, A. (1999). Eye Tracking in Usability Testing: Is It Worthwhile? Workshop on Usability & Eye Tracking, CHI'99.
- Ball, L., Eger, N., Stevens, R., & Dodd, J. (2006). Applying the PEEP Method in Usability Testing. *Interfaces*, 67, 15-19.
- Desurvire, H. M., & Toth, J. (2004). Using heuristics to evaluate the playability of games. CHI '04 extended abstracts on Human factors in computing systems, 1509-1512 .
- El-Nasr, M., & Yan, S. (2006). Visual Attention in 3D Video Games. Proceedings of the 2006 ACM SIGCHI international Conference on Advances in Computer Entertainment Technology (ACE '06), California, June 14 - 16, 2006 .
- Fabricatore, C. M., & Rosas, R. (2002). Playability in Action Video Games: A Qualitative Design Model. *Human Computer Interaction*, 17, 4, 311-368.
- Goldberg, H., & Kotval, X. (1999). Computer Interface Evaluation Using Eye Movements: Methods and Constructs. *International Journal of Industrial Ergonomics*, 24, 631-645.
- Guan, Z., Lee, S., Cuddihy, E., & Ramey, J. (2006). The Validity of the Stimulated Retrospective Think-Aloud Method As Measured by Eye Tracking. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Montréal, Canada, April 22 - 27, 1253-1262 .
- Hansen, J. (1990). The Use of Eye Mark Recordings to Support Verbal Retrospection. *Acta Psychologica*, 76, 1, 31-49.
- Hertzum, M. (2006). Problem Prioritization in Usability Evaluation: From Severity Assessments to Impact on Design. *International Journal of Human-Computer Interaction*, 21, 2, 125-146.
- Hornbæk, K., & Frøkjær, E. (2005). Comparing usability problems and redesign proposals as input to practical systems development. *ACM Conference on Human Factors in Computing Systems*, 391-400.
- Jacob, R., & Karn, K. (2003). Eye Tracking in Human-Computer Interaction and Usability Research: Ready to Deliver the Promises. In J. Hyönä, R. Radach, & H. Deubel, *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*. Elsevier.
- Jeffries, R., Miller, J., Wharton, C., & Uyeda, K. (1991). User interface evaluation in the real world: A comparison of four techniques. *ACM Conference on Human Factors in Computing Systems*, 119-124.
- John, B. E., & Marks, S. J. (1997). Tracking the Effectiveness of Usability Evaluation Methods. *Behaviour & Information Technology*, 16, 188-202.
- Jørgensen, A. (2004). Marrying HCI/Usability and Computer Games: A Preliminary Look. Proceedings of NordiChi '04, October 23-27, 2004, Tampere, Finland, 393-396.
- Lin, T., & Imamiya, A. (2006). Evaluating Usability Based on Multimodal Information: an Empirical Study. Proceedings of the 8th international Conference on Multimodal Interfaces (ICMI '06), Banff, Canada, November 02-04, 364-371.
- Malone, T. (1982). Heuristics for Designing Enjoyable User Interfaces: Lessons From Computer Games. Proceedings of Human Factors in Computer Systems, Gaithersburg, Maryland, 63-68.
- Medlock, M., Wixon, D., Terrano, M., Romero, R., & Fulton, B. (2002). Using the RITE Method to Improve Products; a Definition and a Case Study. Proceedings of Usability Professionals Association (UPA), Orlando, FL .
- Nørgaard, M., & Rau, J. (2007). User Testing in the Combat Zone. Workshop of the International Conference on Advances in Computer Entertainment Technology .
- Poole, A., & Ball, L. (2006). Eye Tracking in Human-Computer Interaction and Usability Research: Current Status and Future Prospects. In C. Ghaoui, *Encyclopedia of Human Computer Interaction*. Idea Group.

Rayner, K. (1998). Eye Movements and Information Processing: 20 Years of Research. *Psychological Bulletin*, 124, 3, 372-422.

Redish, J., Bias, R., Bailey, R., Molich, R., Dumas, R., & Spool, J. (2002). Usability in practice: Formative usability evaluations - Evolution and revolution. *ACM Conference on Human Factors in Computing System*, Minneapolis, Minnesota, 885-890.

Schnipke, S., & Todd, M. (2002). Trials and Tribulations of Using an Eye-Tracking System. *CHI '00*

extended abstracts on Human factors in computing systems, 273-274.

Sears, A. (1997). Heuristic walkthroughs: Finding the problem without the noise. *International Journal of Human-Computer Interaction*, 9, 3, 213-234.

Spool, J. (2006). Eyetracking: Worth The Expense? Retrieved from *User Interface Engineering*: <http://www.uie.com/brainsparks/2006/06/13/eyetracking-worth-the-expense>.

Organizational Challenges to User Research in the Video Game Industry: Overview and Advice⁸

Mie Nørgaard

Copenhagen University
Njalsgade 128-132, bygn. 24
DK-2300 Copenhagen
+45 3532 1446
mien@diku.dk

Janus Rau Sørensen

IO Interactive/Eidos Interactive
Kalvebod Brygge 35-37
DK-1560 Copenhagen
+45 2538 0061
janus.sorensen@ioi.dk

Abstract

In this chapter, we take a look at organizational challenges for 3rd party developers who are interested in implementing and conducting HCI-related user research, such as usability testing, in a game development setting. We discuss the challenges related to justifying the return of investment of user research, formalizing work procedures involving user research, and the building of cross-professional relationships amongst key stakeholders to user research. Furthermore, we also discuss the challenges related to the fact that many games developers are owned or closely affiliated with a publisher. Through the lenses of a questionnaire survey including members from the game industry, we specifically look at the relationship between 3rd party developers and the publisher's marketing department, and investigate how and to which extent these two parties collaborate on user research issues. During the chapter we also present concrete advice on how to tackle the various challenges mentioned.

Introduction

There are many potential rewards for the video

⁸This paper is originally published in Isbister & Schaffer (eds.), 2008, *Game Usability: Advice from the Experts for Advancing the Player Experience*, Morgan Kaufman.

game developer who wants to implement methods to evaluate usability or user experience in the game development process—but also a multitude of challenges. Not only are games complicated pieces of software, the nature of their use is also very different from the use of traditional task-oriented software, which is what most usability evaluation methods are designed for, and this poses a challenge for user researchers in the game industry. Well-known usability measures such as efficiency, effectiveness and satisfaction (such as identified by ISO 9241 – 11) can only partially give a picture of how well a game performs. In fact, one may wonder what ‘efficiency’ in relation to video games actually means, or whether the term makes sense in this context at all (Barr, Noble, & Biddle, 2007; Jørgensen, 2004; Philips, 2006; Bernhaupt, Eckschlager, & Tscheligi, 2007; Pagulayan, Steury, Fulton, & Romero, 2003). But adapting methods or designing new ones are not the only challenges for user research in the game industry. In this chapter we discuss organizational challenges for user research—justifying return on investment, formalizing work procedures, and the building of cross-professional relationships. We further identify a challenge that in some aspects is unique for the game industry; it is a challenge that is connected to the developer-publisher relation and springs from the fact that many game developer studios are either formally owned by or are affiliated with a publisher. In this structural setup, the publisher handles e.g. marketing and distribution, while the developer handles the actual development of the game.

Introducing user research in the form of usability or user experience evaluation at the developer site can potentially create a conflict between the publisher and the developer, because both parties—sometimes simultaneously—conduct user studies. Practically speaking, the user research workers at the developer site will do research with, for instance, a usability focus, while the publisher mainly focuses its user research on marketing issues. At the very least such a situation will require intense coordination between the publisher and the developer site, since they need to agree on for example who the users are, and what the consequences of particular results should be; i.e. if and how to use the results in the development and/or marketing of the game.

This challenge of coordinating marketing and development user research efforts is also present in the software industry, but due to the video game industry’s close historical and structural ties with the toy and entertainment industry, the marketing-development relation and power balance in games development are different from those in the software industry. This poses unique chal-

lenges to implementing user research methods in video game R&D.

In the course of this chapter we will use the term UR champion to describe the person who incorporates—or wishes to incorporate—user research in the development process at the developer site. In terms of job roles in the game industry such a person may belong to level design, QA, management, etc. The goal of the chapter is to discuss the organizational challenges such a person may encounter and to provide tips for how to work around them.

We wish to emphasize that the results and advice presented in the following should not be understood as devious tactics to gain world domination for UR champions at the expense of developers, for instance. Neither should it be understood as an attempt to point out neither developers nor publishers as antagonists—both can be quite positive towards user research. On the contrary, it is the authors’ firm belief that user research should help and enable game developers, as well as publishers and marketing, to develop, market, and sell better games. Accordingly, the challenges and advice presented in the following are aimed at how to manage the organizational aspects of implementing and maintaining a new methodology—to the benefit of all parties.

Three well-known challenges

In (Nørgaard & Rau, 2007) we described and discussed four common challenges in incorporating usability and user experience evaluation in the development of commercial video games: 1) justifying return of investment, 2) developing game specific evaluation methods, 3) formalizing work procedures, and 4) the building of cross-professional relationships. This discussion was based on our experience from working with user research at a large Danish game developer under Eidos Interactive. However, we find that the challenges are of such a general nature that they will be relevant in other organizational settings as well. The challenge concerning the methods for involving users in the development of a game is discussed elsewhere in this volume. Accordingly, this chapter elaborates on the remaining three interconnected challenges—justifying return of investment, formalizing work procedures, and the building of cross-professional relationships—all of which concern intra-organizational aspects of major importance to how successful user research can be introduced in a company developing games.

Finally, we describe a challenge that connects to the relationship between developer and publisher, which in some aspects is unique for the game industry, and thus has not been described elsewhere.

Return of investment

Like in any other industry, the production and design methods that survive are the ones, which add the most to a company's profit. Consequently, a key challenge in convincing management or developers to include user research in the game development process relies on the UR champion's ability to adequately describe the return of investment. Return of investment is not a new theme, and has already been discussed in detail in relation to traditional software development by for example (Karat, 1997; Nielsen & Gilutz, 2003). UR champions everywhere may experience difficulty in persuading the company's management and/or development team to allocate time and money to evaluate an upcoming product's usability. In the game industry, skeptics may argue that many games have done well without much usability evaluation or user research studies, and they will in fact be right. So, why do user research, one may ask. To answer this, we need to consider that traditionally, game designers have developed games to users who—experience and preference-wise—were much like themselves. And with such a well-known group of users, the need for intensive user studies was fairly low. Today, however, players are much more heterogeneous (Bateman & Boon, 2005), and user studies are crucial for the success of a game. As a result, UR champions need to persuade game designers that user studies can provide new insights about the users which can be utilized to improve the design and better target the game to the intended users. UR champions need to produce a set of convincing arguments about both short- and long-term benefits of their work if they are to succeed in convincing management and colleagues in spending time and money on for instance usability and user experience work. Pointing to examples from the game industry which document user research being used successfully in game development is a good first step: If the competitors use a method that seems to give them an edge on the market place, this is in itself a good reason to consider implementing similar methods.

But more than this, the UR champion will also need to point to the reasons why a specific method will have a positive impact on game development. An important factor in being able to make any successful pitch is to make the pitch fit the listener's professional and personal profile—just like a good game must fit the targeted user. This should be kept in mind when attempting to

convince different people or whole departments of the generous returns of user research investments. This means that a UR champion needs to identify the key stakeholders (i.e. the key people that are going to pay for it or whose work will be affected by it), understand what specific returns they are interested in, explain to them what kinds of return they can expect, and relate this to the size of the investment they have to make.

Developers in particular may worry that user research will lead to letting the users (or the UR champion) design the games instead of the developers. On the contrary, user research should support and enable the developers' vision for the game rather than take away responsibility and competence, and this should be communicated clearly to the developers. Furthermore, as project schedules are often very tight on time, a reasonable worry on behalf of the developers is that user research will add more hours and stress to an already heavy workload. Therefore, the UR champion needs to present arguments that user research—although naturally requiring some investment of time—enables the developer to identify necessary design changes much earlier than without user research, and thus saves time in the end. Such an argument fits developers as well as management.

Another persuasive argument is hidden in including developers in the preparation, execution, and analysis of user test sessions. This will demystify user research and help developers understand what user research methods are, what kind of results they can provide, and which questions they might help answer. Because of time constraints, it may not be easy to convince developers to take part in user research. This makes it all the more important to emphasize that the developers' knowledge about the game can be invaluable for the analysis of the research data, which calls for the developers' active participation. Furthermore, from a psychological point of view, developers are more likely to act on evaluation results when they have contributed to creating them (Benton, Kelley, & Liebling, 1972; Schindler, 1998) which makes the involvement of developers in user research even more important.

Whereas developers primarily will be focusing on the production side of the game, management will additionally be interested in how user research can help the company on the market place. The UR champion could therefore seek to document current industry trends—such as a diversifying market with new types of users, escalating production costs etc.—and use these as an argument for user research. A well-supported argument that states that user research can align the game better to the market as well as cut costs, is an ef-

ficient argument that states: We cannot afford not to implement user research if we are to remain competitive.

As a last persuasive factor, it is important that results start rolling in fast after the first user tests, and that these results are both easily communicated, relatively uncontroversial, and easily translated into action points. For instance: A lengthy ethnography-inspired field-work study—although potentially yielding interesting and insightful results—is difficult to validate, hard to understand for non-ethnographers, may require deep (and thus complicated) intervention in the game design, and prolongs the time between investment and return. So, introducing user research through thorough ethnographic studies will make it harder for developers and management to accept user research as adding tangible value to development. Instead, much can be gained by some amount of strategic planning. Initially, the UR champion could keep focus on methodologies—such as basic usability testing—which focus on objective data collection criteria and/or relatively isolated parts of the game. As these methods gain momentum, the UR champion could then start expanding the user research toolkit in order to gradually expose colleagues to other user research methods and train colleagues to think in terms of user experience.

Skeptics may object that it is not the job of the user researcher to pick and choose strategically from the pool of results or tools, and that UR champions have an obligation to present whatever results they uncover, despite any practical or political complications. While this is certainly valid from a purely academic standpoint, we do advise practitioners to at least consider the option of a more pragmatic approach. After all, firing all your artillery and using all your ammunition at level one may not be the best strategy to secure success for user studies in the long run.

Key takeaways:

- Collect real-world examples of successful user research practices in the game industry and share them.
- Tailor return of investment arguments to fit key stakeholders' individual and professional needs and goals.
- Be realistic and choose battles wisely: Start off by implementing user research methods with focus on data and objectivity, as well as a high and reliable success rate. This will enable you to build return of investment-credibility fast and open doors to introducing new user research methods.

Formalized work procedures

A methodology cannot truly prove itself unless it is clearly connected to the relevant development processes it intends to support, which is why a key challenge for UR champions is to create, maintain, and further develop formalized work procedures for user research.

Another critical part of game development—QA testing and other established QA processes—has for a long time been an integrated part of the production process. By now, QA has a relatively well-defined place in the development structure and process, the idealized work-cycle being: QA receives the latest build from the developers, the build is tested in different ways against a set of requirements, and discrepancies are entered into a bug/defect database application. Following this, the developer resolves the bug in the code, commits the code, and makes a new build for the QA department to test. Accordingly, it is relatively clearly defined who has which responsibilities at what stage of the workflow. This means that the bug does not end up in a limbo. Similarly, it is also relatively well-defined when in the development process, QA testing should start, when it should finish, and what it should focus on at which stages in production. UR champions need to ensure similar formalized work procedures for user research.

Of course bugs are different from usability problems, and creating the ideal work-cycle that supports user research during the development of a game may sound easy—but it isn't necessarily, since it requires the involvement of the entire development team.

Thus, after having identified what parts of the game to research—probably central game play features and/or key segments of the game—the UR champion needs to embed the user research efforts into the development of these parts of the game. We use the term 'embed' as opposed to 'add' since it is important that user research doesn't become an add-on method, applied when milestones have already been met. In an Agile/Scrum type of development environment this means for example including user research-criteria for when features are done, and similarly in a waterfall production, including user research in the milestone definitions (Cusumano & Selby, 1997; Schwaber & Beedle, 2001).

One general and practical problem for the UR champion is getting access to playable builds that can be used for user testing. Since it is difficult to pre-order playable builds for a certain date, planning user tests can be quite a hassle. The problem increases if stakeholders understand

user research as a less important activity that is merely added to the development. It is crucial to the quality of the user research that the delivery of builds for user studies is an integrated part of the development schedule. UR champions who struggle with work procedures that impede user research by not including it in the development schedule should make it clear to management that the point is not to add more deadlines to the project, but to create work procedures that support user research so time can be saved in the long run. Otherwise user research will remain an add-on that can be cancelled at convenience. The UR champion could argue that if the company wishes to work seriously with user research, work procedures should reflect that a feature is not done until it is user tested. Accordingly, testable builds need to be available for user research during the development process.

When establishing user research as an integrated part of a development processes, user research practitioners face the challenge of determining to whom the user research feedback should be directed, and who is responsible for carrying out which usability recommendations. This is important if results are to actually be used and re-designs implemented. Game development is often organized in a way where different developers are responsible for different parts of the game, for example animation, character graphics, or AI code. Unfortunately, some user research results simply fall between areas of competence because they involve several functional components of the game. And if convincing a developer to deal with user research issues in his own domain is difficult, convincing him to deal with issues outside of his domain is practically impossible. Thus, one important challenge for a UR champion is to get a clear image of who is responsible for what, and to make sure that any usability issues that do fall between areas of competence are somehow still discussed and handled instead of put on hold or ignored. Also, to catch any unsettled issues in danger of being forgotten the UR champion should be prepared to do an extensive amount of follow-up work.

The issue described above is complicated by the nature of the feedback that user research yields: In contrast to traditional bugs, which primarily focus on clear-cut functional defects in the code or the character models, results from user research are less clear-cut. For example, most results from user research can rarely be considered showstoppers that will leave the game entirely unplayable. Instead, they describe issues that—if resolved—will improve on more intangible aspects of the game such as the overall player experience. That being the case, unless formalized work proce-

dures are in place before results start coming in from user research, there is a great risk of the issues being lost in translation or down-prioritized because they are considered less important than bugs. Ultimately, this may very well mean that usability or user experience issues will end up not being resolved.

Related to this, specifications and best practices on how the UR champion shares his or her results with colleagues are needed to improve user research's impact on the product. Current literature confirms that the means by which results from usability evaluations are presented and communicated to developers are highly determinant for how they are received (Nørgaard & Høegh, 2008; Nørgaard & Hornbæk, 2008a). This will vary from organization to organization, and from team to team, so there will be an element of trial and error and gut feeling connected to this. One way of helping such processes along is to agree on who is responsible for and has the mandate to make decisions about usability priorities, who can instigate user research, to whom the results are handed over, and how these results are handled. It is not the authors' opinion that user research results should automatically warrant a fix as would the discovery of a bug; developers may have good reasons for rejecting a proposed re-design. Nevertheless, there should be a clear work-flow for the handling of user research results and recommendations for re-design. To make this easier, we recommend implementing one method and workflow at a time.

Key takeaways:

- Identify key development components and milestones that user research should connect to.
- Build standardized procedures for user research: Make it an integral part of the development process, not just an add-on that can be dismissed when time is tight.
- Use best practices and gut feeling for which format should be used to share the results. Remember that user research should be supporting the developers' goals and the overall company strategy.
- Start slowly, integrating in tiers or one method at a time.

Nursing cross-professional relations

The successful implementation of user research methods do not only rely on work procedures that support user research, it also relies on the UR champion's ability to form sound cross-pro-

fessional alliances and relationships.

In order to boost their impact on colleagues who regard user research with suspicion or reluctance, UR champions may benefit from forming alliances with those stakeholders who take an interest in user research. Through alliances—or tight cross-professional relations, as we diplomatically call them—with powerful colleagues a UR champion may improve the impact of his or her work tremendously. Thus, the successful UR champion has an eye for strategic planning, lobbying, and for spotting influential colleagues.

However, the relations with influential managers are not the only ones that UR champions need to nurse. Because user research ultimately will have impact on most of the development processes the UR champion needs to develop fruitful relations with a whole range of professionals. For example, because game developers' visions for a game are rarely entirely documented, and because UR champions depend on knowing these visions to understand which challenges in a game are intended and which are actual problems, close cooperation with game developers is important.

Having said that, getting the relevance of user research acknowledged by game developers may be fairly difficult. So, apart from justifying the return of investment, UR champions should also pay close attention to the professional and personal relationships that exist between themselves and other stakeholders in an organization.

When seeking to nurse cross-professional relations, UR champions should pay attention to the fact that different professionals have different aims and job roles, and make an effort to build tight relationships with stakeholders bearing that in mind. Personal relationships are—obviously—also very important because good personal relations help bridge conflicting interests and generally facilitate on-going informal communication, the latter being very helpful from a proactive point of view.

To a critical eye, teaming up with influential colleagues and making alliances might seem a little too Machiavellian. However, the point is not to trick people or to force an opinion upon someone, the point is to build and nurse good relations with colleagues in order to aid the development of successful games.

Key takeaways:

- Think strategically: Do lobby work and strive to make alliances with influential colleagues. Remember, an influential colleague is not always the one with the fancier job title.

- Talk with the game developers and listen to their thoughts and ideas. This may sound trivial, but to succeed with user research you need to understand their visions for the game and you will not if you solely correspond per email.
- Nurse professional and personal relationships continually—not only when you need favors or support. An informal chat once in awhile will help get attention and goodwill when push comes to shove.

Based on our own research and experiences, we have presented some key organizational challenges for UR champions working to introduce user research in the game industry. These challenges are not unlike the challenges that any software company will encounter in the process of maturing its view on usability and the processes for conducting user research. Further discussions of such themes can be found in Helander et al.'s *Handbook of Human Computer Interaction* (Helander, Landauer, & Prabhu, 1997).

In the following, we will discuss one organizational challenge that—in some aspects and to the best of our knowledge—is unique for the role of user research in the development of video games. Accordingly, it is not mentioned in traditional literature on user experience or usability work. This challenge is partly linked to the fact that games development have not sprung from the organizational context of traditional software development, but from the publishing and toy industry. Thus, companies that produce games may organizationally resemble a publishing company more than a producer of traditional task-oriented software, and should be understood in that context, even though basic production issues, such as ensuring usability and a good user experience, on the surface are nearly identical to issues in the software industry.

The publisher and the developer

Background

To understand how and why companies that develop video games are structurally different from the ones that produce ordinary software, we will briefly look at how game developers ended up being related to the publishing industry. Most of the readers will be familiar with many of the points in the following, but we believe that this brief history lesson is important for understanding the organizational context of game development.

The dawn of the commercial video game occurred

in the early seventies—up until then, games were basically programmed by engineers to entertain engineers (Bateman & Boon, 2005; Juul, 2005). But with the emergence of successful coin-up games—such as Atari's PONG (Kline, Dyer-Withford, & De Peuter, 2003) intended for public spaces such as bars and cafeterias—video games showed up on the entertainment industry's radar. Soon, toy and media companies joined the games business contributing with knowledge and experience in production, publishing and distribution.

When Atari's VCS-console was released in 1977 it became a big hit but by 1985 video console game sales had dropped from \$3 billion in the US alone to \$100 million worldwide (Miller, 2005). The reasons for this collapse may be manifold. One reason, which is interesting from an organizational perspective, is that the industry was put together in such a way that developers could create games to whatever console they desired, relatively independent of any publisher. And the outlook to make an easy surplus made many types of companies—even breakfast cereal producers such as Quaker Oats—enter the game of developing games. Unfortunately, this gold-digger mentality flooded the market with poorly designed games, and sales dropped accordingly.

To mend the negative sales statistics the console producers now introduced rigorous screening procedures, e.g. proof of concept and technical requirements, to help them decide which games should be published on their particular gaming console (Kline, Dyer-Withford, & De Peuter, 2003). This made it virtually impossible for independent developers to get onto the console game market without a powerful publisher to get them through the screening process.

From an organizational perspective this may be considered a cornerstone in the relationship between publishers and developers: To get a game onto the market, developers now had to go through a publisher (at least when it comes to AAA console games). In this respect, the video game industry is actually closer to the music industry than the software industry. From the nineties and on the bond between developer and publisher tightened, and today many developer studios are owned by a publishing company that manages the distribution and marketing of a video game.

In terms of games evaluation, such an organizational set-up often entails that the publisher will handle the user-centered evaluation (via marketing methodologies) and the developer most of the technical evaluation (quality assurance) through functional tests or bug-testing (Kline, Dyer-Withford, & De Peuter, 2003). Such a distribution of responsibilities seemingly leaves many critical

decisions about user research in the hands of the publisher's marketing department. This is not a bad thing per se, but what happens when someone decides to evaluate usability or user experience at the developer site?

Based on the assumption that no one likes to give away power or influence we expected that such actions might not be welcomed by the publisher's marketing department and that some rivalry might occur between publisher and developer on that account. At the very least, we assumed there would be an increased need for coordinating the user research efforts at the developer studio and the publisher respectively.

To investigate this and to better understand the reality and challenges for user research at the developer site we conducted an informal survey in the video game industry.

Stories from the field

We invited people from the game industry to answer an online questionnaire focusing on issues such as: How is user research carried out in the particular company, what is its focus, and what is the relationship between the people conducting user research from the developer site and from the publisher's marketing department. The invitation was emailed to 80 recipients from our professional network or randomly selected from the games developer database on www.gamesdevmaps.com. An invitation was also posted on the bulletin board on the International Games Developer Association's web page. Participants were promised anonymity, since our request that they share sensitive information about their user research challenges, might put participants in an awkward position if they were to be identified.

Eleven questionnaires were returned. While this may not be an impressive number in and by itself, we had an equal amount of responses that expressed great interest in the topic and regretted not to have time to participate. Since user researchers and other professionals in the game industry hold myriads of job titles we dare not comment on our sample size or the quality of the answers. However, we find that the answers cover both large cooperations with many well-known titles in the past and smaller ones with less experience. Furthermore, informal 'off the record' conversations with people from the industry confirm the findings.

Participants were offered to comment on our findings and discussion in order to increase the relevance and validity of these. Only one participant provided comments and ideas for improvements.

ID	Type	Full-time employees	Age of company/ games department	Location of HQ	Job title	Experience
A	Developer	17	5	North America	General Manager	7 years
B	Developer	25	9	Middle East	CEO	10 years
C	Developer	80	11	Europe	QA Manager	4 years
D	Developer	7	9	North America	Development studio head	25 years
E	Developer/ Publisher	150	10	Europe	Senior QA lead	8 years
F	Developer	500+	14	North America	User research engineer	3 years
G	Developer	n/a	10	Europe	CEO	10 yeras
H	Developer	30	6	Europe	Project lead/producer	9 years
I	Developer	150+	12	Europe	Development director	4 years
J	Developer/ Publisher	500+	22	Europe	Producer	4 years
K	Developer/ Publisher	500+	14	North America	User researcher	3 years

Table 1: A description of the companies and the participants.

On average, the participants had worked 7.9 years with user research or related work in the game industry. Ten of eleven participants had a background in university studies like history, physics, engineering, computer science, cinema or psychology –though some had never finished their degree. One had other education.

Seven of the participants were 3rd party developer studios, that is, a game developer that works under contract with a publisher for each game. Four participants were publishers/developers or mainly publishers. Table 1 shows a description of the companies and participants.

The responses suggested that user research procedures in the game industry are quite diverse, but that publisher and developer in most cases share the work between them (see Table 2 and Table 3 for details).

We also asked which findings or issues the participants look for in the user research they had knowledge of. Table 4 shows which focus areas were described by participants. UI, game play and concept are the focus areas of most user research. It is interesting to see that developer B, which has very limited cooperation with their publisher, and thus has all responsibilities for user research, also deploys methods with traditional marketing foci

such as market analysis.

Table 5 shows the multitude of methods UR champions use to answer their research questions.

What is apparent about the answers is that most participants had difficulty describing the methods they use. We expect that ‘usability test’ describes some sort of practice related to the think aloud protocol, while ‘playtest’ may mean observing or otherwise monitoring users play. Thus, ‘conducting playtests’ may be the same as ‘observing play sessions’. If this is true, observing users play the game is the most commonly used method deployed and in fact the only method used by some of the participants. The lack of clarity in terms of describing the methods used to conduct user research may be because participants were not familiar with research terminology or simply because of a lack of generally agreed upon naming conventions. However, we are more prone to explain it with user research practice being improvised and hardly ever formalized or put into system. Notable exceptions are C and K. They specifically mentioned the aim to triangulate methods and combine qualitative and quantitative methods in order to obtain both objective and subjective data. As discussed earlier we urge UR champions to formalize their procedures, describe the methods they use, and which questions

Who conducts user research?	ID
Developer site	BK
Mostly developer, but some at the publisher site	CGI
Shared equally between developer and publisher	DEHJ
Publisher site	F
Mostly publisher, but some at the developer site	A

Table 2: The distribution of practical user research.

Who takes the initiative to user research?	ID
Developer	BCGJK
Publisher	EA
Both	FHI
n/a	D

Table 3: The distribution of user research initiative.

these particular methods can help them answer. Such work will yield the most reliable results and thus boost the credibility of the user research. Without this formalization work done, the results of the user research are more vulnerable to unvalidated ‘common sense’ objections.

Since we wanted to investigate the relationship between publisher/marketing and developer, we asked participants if and how user research results were shared. Five participants answered that they hardly have any communication with the publisher’s marketing department about user research results. One developer mentioned being very interested in getting data from the marketing department and another that the publisher was unlikely to be interested in the developer’s user research.

With regards to the sharing of results, developer B mentioned how they mostly communicate early user research results to the publisher as an attempt to make them ‘join the adventure’. Such a sharing of results thus seems mostly motivated by the wish to land a contract. Along the same lines developer D described how both developer and publisher manipulate their user research results before sharing them with the other party. J described how user research results from the publisher are shared with the developer studio and vice versa, but also suggested that not all results were to be shared with everyone. Related to this, K described how user research results were kept from the marketing department on purpose—the rationale behind this was that marketing tended to misinterpret preliminary results and base marketing and approval decisions on them, thus effectively causing development teams to not want to work with the user researchers. Similarly, K describes how development teams only listened

to marketing user research results (such as focus groups) so as to please marketing with the ultimate goal of ensuring a marketing budget for the game; not really to make any changes in the game design based on the results.

These answers may suggest that some of the communication and relationship between developer and publisher is not primed to actually facilitate better collaboration on the shared goal: Making a good and successful game. Rather, they seem to suggest that developers do not always consider the publisher a friendly colleague but rather a partner that needs to be maneuvered to fit the developer’s goals. And that the same goes for the publisher. On the other hand, it is only to be expected that developer studios and publishers see the world from different perspectives, and therefore it is no surprise that they have different goals for user research. Nevertheless, this points directly to a need to coordinate user research efforts.

Participants described how the relationship between developer and publisher isn’t all roses. The lack of knowledge about what goes on on the other side of the fence impedes and slows down production and coordination. This is suggested to affect creativity and probably, in particular unfortunate cases, ultimately sales.

One developer explained how the very nature of being a 3rd party developer means that the publisher has the most rights to the game. And this is suggested to cause some imbalance in the relationship. Conversely, a publisher described being helplessly dependent on the developer to implement the changes that arise from for example focus tests. This is also suggested to cause unevenness in the relationship, mainly because the timing of user research is of huge importance

Focus area	ID
Acceptance of concept	BDEFGHIJ
Problems with game play	AFGHIJK
Do users understand UI?	ADGK
How game is perceived in different markets	EIJ
Market and competitor analysis	BIJ
Fun	CDK
Quantitative measures (e.g. number of times died/preferred weapon/playtime)	CK
How well does game correspond with the brand and its values	E
Estimation of sales numbers	B

Table 4: The focus areas for producers’ user studies.

Method	ID
Filmed or observed play sessions	EFGHIJK
‘Playtest’ (e.g. with participation of friends)	ACDIJK
Interviews with target users	CEHIJK
Focus groups	CIJ
‘Surveys’	FK
Usability tests	FK
Questionnaires	CK
Data-logging	CK
Rudimentary testing with users through web site	B

Table 5: The methods used to conduct user research.

to the relationship and that fights are bound to break out if user research results are forced into the development at too late a stage in the development process. As an example, changes that will require large investments are mentioned as an issue giving rise to severe challenges for publisher-developer cooperation.

In all fairness it should be emphasized that three participants from publishing or publishing/developer companies generally were very satisfied with their communication with the developers and the planning of user research. However, the three companies are fairly large and experienced publishers, and this may explain why they pay attention to and enjoy success implementing effective work procedures around user research. Since the developers in the study seemed more concerned about the state of the communication and collaboration between developer and publisher, we do speculate whether developers in general feel more insecure or unsatisfied simply because they are the less powerful party of the two.

One publisher explained that user research results rarely get completely ignored, and that developers often have a good reason for putting results on hold. Such a comment shows a rare and valuable understanding for colleagues’ points of view, and confirms that much is accomplished by

trying to understand colleagues’ motivations and goals. This supports the importance of building and nursing the cross-professional and personal relationships.

Another publisher specified how not being able to communicate directly with a 3rd party developer was a huge challenge. Direct, informal and often occurring communication was claimed to be crucial to the publisher, who needs to be up to date with the development process and recent game builds. ‘Getting to know each other’ secures that colleagues are accessible, that they listen, and that they are honest in their communication’, the publisher suggested, emphasizing the value of personal relationships. In this context, the building of cross-professional relationship should be seen both in an intra-organizational and trans-organizational context.

A developer explained how there seems to be a semantic gap between development and marketing: that the developer seemingly has difficulty understanding what exactly marketing does and vice versa. This was confirmed by other participants that mentioned a need for creating a better understanding for user research methods on each side. Such efforts should provide greater transparency for what research is being done in each camp and what questions it is supposed to an-

swer. Related to this, one participant suggests that the marketing department needs a higher level of methodological rigor in their user research and an increased awareness of what methods can assess what questions: Focus groups should not be used to validate design, but instead function as a point of departure for brainstorming design ideas.

Reflections on the results

Involving users successfully in systems development is never an easy feat. Numerous accounts on the difficulties of this task have been given in relation to the development of office-ware, web applications etc., see for example (Gould, Boies, & Ukelson, 1997). Some of this work specifically point to how organizational issues (Iivari, 2006) and the relationship among job roles may impede the impact of usability on design (Furniss, Blandford, & Curzon, 2007; Gulliksen, Boivie, & Göransson, 2006; Nørgaard & Hornbæk, 2008b).

Grudin and Markus (Grudin & Markus, 1997) described how contract development has often ended up creating substantial barriers between developers and users, and how the separation between developers and users—as in cases where for instance a marketing department monopolizes user contact—presents a major organizational obstacle for design in contract development. Gould and Lewis' discuss similar issues in their classical paper on key principles of design (Gould & Lewis, 1985). Other records describe how marketing departments are reluctant to share the opportunity to get in first hand contact with users, or perhaps forbid other departments to do it all together (Grudin, 1991; Frøkjær, 1987).

Based on some of the anecdotes we have heard in the game industry, we wondered if the same was true for the relationship between a publisher's marketing department and a 3rd party game developer. While our study clearly contains examples of it, the results are not univocal: The horror story frequency in the answers was in fact very low. However, some of the results as well as informal communications we have had with participants suggested that perhaps the developer-publisher relationship is a bit more complicated than described by the answers we received. We have come across anecdotes that imply that it may be a challenge for some UR champions to get to do user research on the developer site at all. Some of the developers in this study have also described their relationship to the publisher's marketing department as being a bit tense, and it was suggested that user research results sometimes were kept away purposefully from the marketing de-

partment. Some also implied that the publisher's marketing department considers a video game the publisher's property, and behaves jealously if attempts are made from the developer site to take control of user research. In this way, the historical structures that lie behind the publisher/developer relationship, where the publisher often decides the fate of the games, potentially makes it harder for the UR champion to implement user research at the developer site, since prior experience with user research methods such as focus groups (performed by marketing) in some cases has created mistrust against user research methods in general.

Once again, the overall challenge as we see it, is that the developer's UR champion and the publisher's marketing department both work with user research, and determine which methods should be used for what insights. However, even though they may share the goal to produce a good and successful game, their focus areas, methods, challenges, and timing are different. This should be crystal clear to anyone who does user research, but unfortunately it is not always.

The developer is often basically interested in how the game works, how fun it is, how difficult it is, and so on. The publisher's marketing department, on the other hand, is basically interested in how the game fits the target audience and the market in general, how it is presented to potential buyers, and so on. Before commencing on a new game marketing may thus choose the customer segment, conduct focus group interviews with potential users, and perform other surveys related to users. When the game is close to being finished, it will then conduct more user tests. To reach its goals the marketing department will also involve users when creating a marketing strategy or settling on a name for the game.

But, while the publisher's marketing department may investigate issues that are closely related to usability and user experience it does not conduct user research in the way it is traditionally understood in the software development or usability consultancy industries. For example, the experience-centered evaluation methods that marketing deploy are often traditional methods for evaluating consumer goods. These include focus groups, market surveys, systematical collection of sales data and pre-production questionnaires. Accordingly, when the publisher's marketing department assures the developer that user research is carried out, it may imply the use of traditional marketing methods before and after production rather than through the qualitative HCI-methods deployed by traditional software producers. And UR champions need to make sure that marketing

will not dismiss any user research on account that they have already done it—because most likely they haven't. Equally important is the coordination of results as they roll in: If user studies at the development site unveil new and crucial knowledge about the target users, then this knowledge needs to be disseminated to the marketing department, since this knowledge could be interesting enough to have an impact on the marketing strategy. Conversely, if marketing methods deployed at the publisher site show new preference patterns from the target audience, this needs to be communicated to the developer site. Of course not all results are crucial enough to warrant design or marketing plan changes, but nevertheless formalized coordination processes need to be in place, preferably in an atmosphere of trust and not mistrust. Making this happen will in some cases require a significant amount of 'marriage counseling' or even a restructuring of the relationship between publisher and developer.

Now, our study suggests that borders between responsibilities and focus areas are not always clearly separated according to whether one is a developer or a publisher. Developers sometimes do market research and publishers sometimes conduct game play or feature-centered user research with an HCI focus. However, we still find that a great responsibility lies with the UR champion at the developer site in making it clear that when developing a video game, user research is not only traditional marketing research. It is also user research as it is understood in an HCI context. Apart from a different methodology this means conducting user research in close contact with the potential users of the game and the people developing the game. This is where we see a great opportunity for the UR champion to spearhead the linking together of developer and publisher, and to create the best possibilities for relevant user research.

Our results also suggest, what was also intuitively expected, that the successful coordination of user research efforts between developer and publisher to a large extent depends on organizational proximity, that is: The closer the organizational ties between development studio and publisher (for instance in the case of a publisher-owned development studio), the better the flow of information. This is not to say that a healthy relationship will always be present, as some of our results also show, but at least very good organizational preconditions for creating and sustaining trust and common goals are present. Conversely, the further apart a publisher and a development studio are—organizationally speaking—the bigger the challenges for coordinating user research efforts. In any event, it is the authors' clear recommen-

dation that every effort should be made to build trust between developer and publisher, so user research efforts can be coordinated, since failure to do so will entail a high risk of incommensurable views on the user and the game's future impact on the marketplace.

Key takeaways:

- Be aware that 'user research' and 'user research' do not mean the same thing in terms of marketing and development.
- Work to build trust between publisher and developer, for instance by sharing your thoughts on methods and research questions with marketing.
- HCI-related user research should be done in close contact with potential users and the people developing the game.
- The bigger the distance between developer and publisher, the bigger the challenge of coordinating user research work.

Conclusion

In many ways the challenges UR champions encounter when striving to do user research in the game industry are similar to the ones they would encounter if they were developing traditional office-ware or other task-oriented systems. Such challenges include justifying that the investment made in user research, creating company work procedures that support user research, and developing professional and personal alliances with key stakeholders. However, since many 3rd party developers are either owned or tightly affiliated with a publisher, some organizational challenges for user research in the game industry are—in some aspects—quite unique.

Our survey amongst eleven developers/publishers from the game industry suggest that a close cooperation between a 3rd party developer and the publisher's marketing department is crucial, but also that UR champions need to pay attention to some of the obvious dangers of doing user research in two separated camps. One danger is that the publisher's marketing department confuses marketing related user research with HCI-related user research and—thinking it is all the same thing—miss the HCI-perspective on a game, and accordingly ignores great opportunities to link the development of a game close to potential users and to the people who develop the game. Another danger is inefficient work procedures caused by the geographical distance and perhaps

also mismatching ideas about how, when and by whom user research should be carried out. There will be variations as to how the described challenges will manifest themselves in different organizational settings, but we expect the basic mechanisms behind the challenges to be present in most game development settings.

Since the success of user research at the developer site ultimately rests on the UR champion's shoulders, we have presented some key take-aways that we believe will help anyone who is interested in conducting this work navigate through the most common organizational challenges.

Acknowledgements

We wish to thank those who took the time to participate in our study and their willingness to share their thoughts. Also, we thank the editors of this book, and Erik Frøkjær for valuable discussions and comments.

References

- Barr, P., Noble, J., & Biddle, R. (2007). Video Game Values: Human-Computer Interaction and Games. *Interacting with Computers*, 19, 180-195.
- Bateman, C., & Boon, R. (2005). 21st Century Game Design. Rockland, MA, USA, Charles River Media.
- Benton, A. A., Kelley, H. H., & Liebling, B. (1972). Effects of Extremity of Offers and Concession Rate on the Outcomes of Bargaining. *Journal of Personality and Social Psychology*, 24, 73-83.
- Bernhaupt, R., Eckschlagler, M., & Tscheligi, M. (2007). Methods for evaluating games: how to measure usability and user experience in games? Proceedings of the International Conference on Advances in Computer Entertainment Technology (ACE'07).
- Cusumano, M., & Selby, R. (1997). How Microsoft Builds Software. *Communications of the ACM*, 40, (6).
- Frøkjær, E. (1987). Styringsproblemer i det offentlige edb-anvendelse. *Politica, Tidsskrift for Politisk Videnskab*, 19, 1, 31-56, (In Danish).
- Furniss, D., Blandford, A., & Curzon, P. (2007). Usability Work in Professional Website Design: Insights From Practitioners' Perspectives. In E. Law, E. Hvannberg, & G. Cockton, *Maturing Usability: Quality in Software, Interaction and Value*, 144-167, Springer London.
- Gould, J. D., & Lewis, C. (1985). Designing for Usability: Key Principles and What Designers Think. *Communications of the ACM*, 28 (3), 300-311.
- Gould, J., Boies, S., & Ukelson, J. (1997). How to design usable systems. In M. Helander, T. Landauer, & P. Prasad, *Handbook of Human-Computer Interaction*. Elsevier Science.
- Grudin, J. (1991). Interactive Systems: Bridging the Gaps Between Developers and Users. *Computer, April issue*, 59-69.
- Grudin, J., & Markus, M. L. (1997). Organizational Issues in Development and Implementation of Interactive Systems. In M. G. Helander, T. K. Landauer, & P. V. Prabhu, *Handbook of Human-Computer Interaction* (second ed., Vol. 1, pp. 1457-1474). Amsterdam, Elsevier Science B.V.
- Gulliksen, J., Boivie, I., & Göransson, B. (2006). Usability Professionals - Current Practices and Future Development. *Interacting with Computers*, 18, 568-600.
- Helander, M., Landauer, T. K., & Prabhu, P. V. (1997). *Handbook of Human Computer Interaction*. Elsevier Science.
- Iivari, N. (2006). 'Representing the User' in Software Development - a Cultural Analysis of Usability Work in the Product Development Context. *Interacting with Computers*, 18, 635-664.
- Jørgensen, A. (2004). Marrying HCI/Usability and Computer Games: A Preliminary Look. Proceedings of NordiChi '04, October 23.-27. 2004, Tampere, Finland, 393-396.
- Juul, J. (2005). *Half-Real: Video Games Between Real Rules and Fictional Worlds*. Cambridge, MA, USA, MIT Press.
- Karat, C. (1997). Cost-Justifying Usability Engineering in the Software Life Cycle. In M. Helander, T. K. Landauer, & P. V. Prabhu, *Handbook of Human-Computer Interaction*. Elsevier Science.
- Kline, S., Dyer-Witthof, N., & De Peuter, G. (2003). *Digital Play: The Interaction of Technology, Culture, and Marketing*. Montreal, McGill-Queen's University Press.
- Miller, M. A. (2005, April 1st). History of Home Video Game Consoles. InformIT, <http://www.informit.com/articles/article.aspx?p=378141&seqNum=3&rl=1>.
- Nielsen, J., & Gilutz, S. (2003). Usability Return on Investment. Nielsen Norman Group.
- Nørgaard, M., & Høegh, R. T. (2008). Evaluating Usability - Using Rhetorical Models to Improve the Persuasiveness of Usability Feedback. Proceedings of the 7th ACM Conference on Designing Interactive Systems (DIS2008).

- Nørgaard, M., & Hornbæk, K. (2008a). Exploring the Value of Usability Feedback Formats. *International Journal of Human-Computer Interaction*, in press.
- Nørgaard, M., & Hornbæk, K. (2008b). Working Together to Improve Usability: Challenges and Best Practices. Retrieved from Technical report from Copenhagen University Dept. of Computer Science,; <http://www.diku.dk/publikationer/tekniske.rapporter/rapporter/08-01.pdf>
- Nørgaard, M., & Rau, J. (2007). User Testing in the Combat Zone. Workshop on Methods for Evaluating Games - How to measure Usability and User Experience in Games, The International Conference on Advances in Computer Entertainment Technology (ACE'07), June 13-15, 2007, Salzburg, Austria.
- Pagulayan, R. J., Steury, K., Fulton, B., & Romero, R. (2003). Designing for Fun: User-testing Case Studies. In M. Blythe, K. Overbeeke, & A. Monk, *Funology: From Usability to Enjoyment*, 137-151, Springer.
- Philips, B. (2006). Talking About Games Experiences: A View From the Trenches. *Interactions*, 13, 5, 22-23.
- Schindler, R. M. (1998). Consequences of Perceiving Oneself As Responsible for Obtaining a Discount. *Journal of Consumer Psychology*, 7, 371-392.
- Schwaber, K., & Beedle, M. (2001). *Agile Software Development with Scrum*. Upper Saddle River, NJ, USA, Prentice Hall.