

Examensarbete

**Turning Usability Goals into  
Measurable Objectives**

av

**Martin Karlsson**

LiTH-IDA-Ex-Ing-02/25

2002-10-17

Examensarbete

# **Turning Usability Goals into Measurable Objectives**

av

**Martin Karlsson**

LiTH-IDA-Ex-Ing-02/25

2002-10-17

Handledare: Ivan Rankin, Andreas Björklind

Examinator: Niklas Hallberg

---

## **Abstract**

Usability is typically measured relative to a user's performance on a given set of tasks. A number of different goals are used as a foundation for usability testing. The aim of this study was to discover whether the usability goals could be turned into measurable objectives.

The most common measures used in usability testing are time, error rate and the user's subjective satisfaction. In this study, a conversion from usability goals to objectives using these measures was performed. User tests were carried out. The test suites were composed of interviews and cognitive walk-throughs of prototypes. The study was conducted with five users. While this is a small number for performing a quantitative statistical study, it has been shown that useful results can be achieved with a small number of subjects.

The results of the study show that usability goals can be turned into measurable objectives. These objectives can be used to give a pointer in qualitative research, but they should not be used only to achieve statistical validity.

## **Sammanfattning**

Användbarhet mäts vanligen genom att kontrollera användarens prestation vid utförande av givna uppgifter. Ett antal olika mål används som grund för användbarhetstestning. Målet med denna studie var att undersöka om användbarhetsmål kunde omvandlas till mätbara mål.

De vanligaste måtten som används inom användbarhetstestning är tid, feluppskattning och användarens subjektiva upplevelse. I den här studien omvandlades användbarhetsmålen till mål som använder dessa mått. Användbarhetstesten var uppbyggda av intervjuer och kognitiva walk-throughs av prototyper. Studien utfördes med fem användare. Trots att detta är ett lågt antal användare om man jämför med kvantitativa studier, så har det visat sig att användbara resultat kan erhållas med endast fem användare.

Resultatet av studien visar att användbarhetsmål kan omvandlas till mätbara mål. Dessa mål kan användas för att ge en indikation i kvalitativ forskning, men bör inte användas bara för att uppnå statistisk validitet.

---

---

## **Acknowledgements**

I would like to thank my instructors Andreas Björklind and Ivan Rankin, and my examiner Niklas Hallberg, for their input in this work. I would also like to thank Annika Holmqvist, Johanna Daag and the people at Xpedio Linköping Ubiquitous Research Center for valuable input in different ways.

Finally I would like to thank friends and family for great support, and all people who have taken part in my user tests.

This report ends my academic years of learning, learning and learning. But I do not want to see it as an ending, since I believe this type of research in this particular area is the most fun anyone can have and still get paid.

Martin Karlsson, Norrköping  
Thursday, 17 October 2002

---

---

## Table of contents

<b>1</b>	<b>INTRODUCTION .....</b>	<b>1</b>
1.1	PURPOSE .....	1
1.2	OBJECTIVES OF THIS STUDY .....	1
1.3	TARGET AUDIENCE .....	1
1.4	DELIMITATION .....	2
1.5	DOCUMENT OVERVIEW .....	2
<b>2</b>	<b>WHAT IS USABILITY?.....</b>	<b>3</b>
2.1	USABILITY GOALS .....	3
2.2	USABILITY METRICS .....	4
2.3	HOW TO MEASURE USABILITY .....	5
<b>3</b>	<b>QUANTITATIVE AND QUALITATIVE RESEARCH.....</b>	<b>7</b>
3.1	TRUSTWORTHINESS .....	7
<b>4</b>	<b>METHOD.....</b>	<b>9</b>
4.1	RESEARCH TECHNIQUES.....	9
4.2	USERS .....	14
4.3	TESTING THE GOALS IN PRACTICE .....	16
<b>5</b>	<b>RESULTS.....</b>	<b>21</b>
5.1	RESULTS FROM THE STUDY .....	21
<b>6</b>	<b>DISCUSSION .....</b>	<b>23</b>
6.1	SUMMARY .....	27
<b>7</b>	<b>FUTURE WORK .....</b>	<b>28</b>
<b>8</b>	<b>REFERENCES.....</b>	<b>29</b>
	APPENDIX A: PILOT TEST QUESTIONNAIRE .....	31
	APPENDIX B: TEST SUITE 1 .....	33
	APPENDIX C: TEST SUITE 2.....	35
	APPENDIX D: INTERVIEW QUESTIONS TABLE .....	37
	APPENDIX E: LOW-FIDELITY PROTOTYPES.....	39
	APPENDIX G: TEST RESULT SUMMARY.....	44

---

This page intentionally left blank.

# 1 Introduction

Usability is the measure of the quality of a user's experience when interacting with a product or system. The usability of a product or a system is hard to measure, but one way to focus a study or a process on usability is to use what are called measurable usability goals. These are relatively simple concepts that make it easier to create tools or systems that are easy to use.

A web portal is a dynamic site on the Internet that provides a comprehensive entry point for a huge array of resources and services. A portal usually contains articles about different subjects. The word dynamic implies that the site content changes all the time. Hence, there is a need for an easy way to update and maintain the site content. A content editor is needed.

## 1.1 Purpose

The intent of this report is to discover whether usability goals can be turned into measurable objectives and whether they are relevant for a practical usability process.

## 1.2 Objectives of this study

The main objective of this study was to perform an investigation of a specific set of usability goals to see to what extent they are relevant in a practical usability process and the extent to which these usability goals are measurable.

In support of the main objective, it was necessary to create a requirements specification for a content editor, to create a usable interface from these requirements and to implement a functional and usable content editor based on the analysis of the usability goals.

## 1.3 Target audience

This report is primarily intended for readers interested in usability. Some knowledge of quantitative research methods in social science is assumed.

### 1.4 Delimitation

The content editor only serves as background to this report. This is mentioned only when required for the purpose of clarification.

A lot of different techniques could be used in usability testing, but in this study only the commonly used techniques are used.

### 1.5 Document overview

This document provides a brief background to what usability is, common usability goals and metrics, in section 2. Section 3 explains the method and techniques used in the study. The participants are also considered here. Section 4 deals with the results of the study that are discussed in section 5.

#### Note 1

From now on in the report, the content editor application is simply mentioned as *the tool*, to simplify.

#### Note 2

This work was performed at Xpedio AB in Linköping. Xpedio is a Stockholm-based company that, amongst other things, offers a Java-based platform for the development, deployment and operation of a dynamic portal server solution. Simple types of content such as news, alert services and so forth may be stored in the adjoining content database. The portal also handles generic types of subscription services for news and alerts. This makes it possible to create a dynamic portal where users can read articles and even import them to their mobile phones including a variety of other platforms like WAP and Japanese/Dutch I-mode.

## 2 What is usability?

To say that a product is usable does not tell us very much. When thinking about a product, we hardly ever think about the main feature of the product, only that it is convenient to use and it has these and those features that will help you. If we consider a car as the product, we won't think that the car might not move at all when we use it. This is because the main reason that people use cars is to move around. If a car won't move, it is not a car. Holmlid<sup>1</sup> tells us that we focus on other things, like the colour of the car, whether it is cheap to maintain and whether it is pleasant to drive.

If we apply these rules to computer programs, several different features have to be considered because of all the different needs that the users might have. For instance, certain features in a user interface, good error prevention and flexibility are appreciated. Then you tend to forget about the feeling you get when using the product, as was mentioned above.

### 2.1 Usability Goals

A way to measure usability is to use some sort of usability goals. These are qualitative. This means that they can be categorized, but will not reduce to numerical measurements (Faulkner, 2000). Kindborg (1999) says that the goals are derived from values the users bring into the task using the product. Is the product heavy or light? Hot or cold? Would you be happy if you got the product as a birthday present? Would you like to give it to someone as a birthday present? To say that a computer program should be easy to learn is an example of a usability goal. But a goal should be based on functional features that are free from look-and-feel constraints. "*The car should be red because it feels good*" is not a good usability goal, while "*The car should have a colour that produces greater satisfaction*" is.

When you create a usability goal, you usually list the issues that you find important and try to generalize them. For instance, a physician using a diagnosis application want to be sure she can get all the patient information she needs with ease so that she can make a good diagnosis. The corresponding usability goal will probably look like this: Users are satisfied that the patient information is easy to access and that it is organized so that it doesn't interfere with the patient's

---

<sup>1</sup> Stefan Holmlid, by email 10 Feb 2002

treatment, or better yet, if the information is organized so that it contributes beneficially to the treatment. (Hackos & Redish, 1998)

### 2.1.1 REAL

A perhaps easier way to measure usability is to use a predefined set of general usability goals. The International Organization for Standardization (ISO) defines usability as “... *the effectiveness, efficiency and satisfaction with which specified users could achieve specified goals in particular environments...*” (Faulkner, 2000, p. 114) This provides a good starting point, but because of the abstract terminology there are far too many problems with this statement, so a more specific solution is needed.

Löwgren (1993) proposes that usability is the result of four specific usability goals, namely *Relevance, Efficiency, Attitude* and *Learnability* (REAL)

- The relevance of a system is how well it serves the user’s needs.
- The efficiency defines how well the users can perform a task using the product.
- Attitude is the user’s satisfaction with the product.
- Learnability denotes how quickly the user can grasp how the product works and how well the user can remember the skills over time.

A lot of different goals are often used to conduct usability studies. Löwgren's four goals appear to be a suitable and well-defined subset and for this reason REAL has been used in this study.

## 2.2 Usability metrics

Usability can be measured in more precise ways, but it rarely is. Qualitative insights when using usability goals are often all that is needed for practical design projects. Nielsen (2001) claims that measuring usability often costs a lot more than doing qualitative studies, but usability metrics let you track progress in and between software releases and assess your competitive position. Thus, a conversion of the usability goals to usability metrics comes in handy.

Measurable objectives can often be more difficult to create than goals (Hackos & Redish, 1998). The reason for this is that they are dependent on the user’s view of what the system should be able to do. For instance, the users may think that they should be able to install the product without trouble. But when they are studied installing

the product, many of them tend to need assistance. The usability goal and its corresponding objective are shown in Table 1. The measurement provides a way to determine whether an acceptable level of usability has been achieved.

Table 1. From Hackos & Redish. 1998, p. 349.

Usability Goal	Measurable Objective
Users will find the installation process understandable and easily follow it step-by-step to achieve a successful installation.	No more than 10% of users will call customer support for help to install the product.

### 2.3 How to measure usability

It is easy to specify usability goals and measurable objective, but hard to collect them. Eason (1984) states that systems are usable if they are in fact used in practice, but how can you measure usability until you install the system? Löwgren (1993) argues that you should base your measurements on his definition of usability goals (REAL).

Nielsen (2001) explains that usability is typically measured relative to a user's performance on a given set of tasks. The most common measures are *time*, *error rate* and the user's subjective *satisfaction* rated on some kind of scale.

Time-based usability metrics are harder to turn into measurable objectives than non-time-based, says Nielsen (2000:2). For example, how long should it take a user to book an airline ticket to London? But measuring time is unproblematic. Faulkner (2000) suggests that measuring how much time a task takes to complete could be used to judge the efficiency of a system. Time is simply a measurement from the beginning to the end of whatever action is being observed.

To measure error rate is not as easy as to measure time. What need to be measured are mistakes, and not slips. A slip is when you put your socks in the fridge instead of the washing machine. On the other hand, a mistake is when you really thought that the socks would get clean in the fridge, ie. a slip is an error made unintentionally, while a mistake is an erroneous belief that one is doing something correctly (Lewis & Norman, 1986). When measuring the error rate it is a good idea to add a severity scale, says P. Jordan (in Faulkner, 2000). A suggested scale

is a four-level one including minor errors, major errors, fatal errors and catastrophic errors. The benefit of measuring error rate, states Nielsen (2000:2) is that it can easily be collected while running a think-aloud study (see section 3.3.4) where the users are asked to verbalize their thoughts and say what they are thinking at each step of the way. With this method, you can continue to collect qualitative insights while you collect a formal usability metric.

Combining these opinions and Löwgren's REAL goals with Hackos & Redish's suggestions (Section 2.3) results in Table 2.

*Table 2.* The REAL usability goals turned into measurable objectives.

<b>Usability goal</b>	<b>Qualitative objective</b>	<b>Measurable objective</b>
Relevance	The tool should be able to perform the tasks that the user wants to perform.	Scale
Efficiency	When performing a task, the user should accomplish the task in the least time and with as little effort as possible.	Time, error rating and subjective satisfaction.
Attitude	The tool should be a pleasure to use.	Scale
Learnability	It should be easy to learn how to use the tool.	Time

### 3 Quantitative and qualitative research

Research methods can be differentiated by treating data in two different ways. Quantitative research methods were originally developed in the natural sciences to study natural phenomena. Quantitative research uses surveys and polls to deliver results that can be used for mathematical modelling. It is utilized when drawing conclusions from a population. While it can cover many people quickly, it does not provide a means to follow-up with questions to explain unexpected answers. Qualitative research methods are designed to help researchers understand people and the social and cultural contexts within which they live. This includes fewer, more in-depth interviews to give findings an added dimension, enlightening the "whys" behind the respondents' answers (Breakwell et al., 2000).

#### 3.1 Trustworthiness

By looking at what Trochim (2001) calls trustworthiness, we can assure that the results of this study are valid. In qualitative research, Trochim states that the trustworthiness of the study, which is usually called validity in quantitative research, can be addressed using four criteria: credibility, transferability, dependability and confirmability.

- *Credibility* corresponds to internal validity in quantitative research. It aims to establish that results in qualitative research are believable in the eyes of the participants. Hence, the credibility criterion is meant to ensure that the subject of the enquiry has been correctly described.
- *Transferability* corresponds to external validity in quantitative research. This criterion queries just to what extent these findings can be generalized.
- *Dependability* is concerned with whether the process of the research produces the same results on different occasion, independent of time, researcher and method. This is fairly the same as reliability in quantitative research.
- *Confirmability* corresponds to objectivity in quantitative research. This criterion addresses the issue of researcher bias and other distortions.

## Turning Usability Goals into Measurable Objectives

---

If your task is to determine how many people like an idea, to measure the size of a market or to prepare a volume estimate, quantitative research is what you need. If instead, you are trying to improve a product or service, identify different market segments or develop a persuasive advertising or sales message, then qualitative research is the way to go (Breakwell et al., 2000). In typical usability studies qualitative methods are used. This is also the fact in this study even though it is about turning usability goals into measurable objectives.

## 4 Method

This section deals with the different techniques used when conducting a usability study, and the given approach of this study.

### 4.1 Research techniques

When designing a completely new product from scratch, Hackos & Redish (1998) suggest that several different techniques should be used. The techniques will help ensuring that the correct task data will emerge from the study. They suggest confronting the user with many techniques to obtain as many aspects as possible very early in the design phase. This is of course not a cost-effective way to perform a study, though it is probably very much needed. Nielsen (1994:1) has defined something he calls Guerrilla HCI or discount usability engineering. This method is based on the use of three techniques, namely scenarios; think aloud protocols and heuristic evaluation. These techniques are explained later in this section.

In this study we will use a simple form of usability study called formative evaluation that, in thought and practice, includes Nielsen's discount usability engineering. Formative evaluation is a method of judging the worth of a program while the program activities are forming or happening. It is mainly based on qualitative social science research. The alternative is called a summative evaluation, which is a method for evaluating a finished and complete system. For this reason, the study is heavily based on the users' assessments of the prototypes discussed in section 3.3.5 (Faulkner, 2000).

The sociologist, Fredrik Engelstad, declares that qualitative research methods are used to differentiate apples and pears, while quantitative research methods defines just how many apples and pears there are. He says that these two method types shape each other, that they should be combined to gain the most from a study. This combination is called method triangulation. To combine different methods from different method types gives a broader spectrum and a safer ground for the interpretation of the results (Repstad, 1999).

It is also essential to combine some qualitative methods, for example observation and interview, to see what the user is doing and whether it fits with what the users think, says Repstad (1999). The following sections in this document deal with these ideas and the different methods used in the study.

### **4.1.1 Questionnaires**

A questionnaire is a method for the elicitation and recording of data. It is a device that starts a process of discovery in the mind of the respondent while recording this onto a permanent medium. The advantage of using a usability-based questionnaire is that it provides feedback from the point of view of the user. But designing the perfect questionnaire is probably impossible. It is usually designed to fit a number of different situations (due to cost reasons). As a consequence, it cannot say what is right or wrong with the artefact that is being tested. However, a well-designed questionnaire can get you close enough. Another problem with questionnaires is that they only tell you the reaction of the user as the user experiences the situation instead of what really happens (Kirakowski, 2000).

To get an even lower error rate in the questionnaire, the questions should be in an open-ended-format. This means that the respondents are asked to write down their response to a question in any terms that they see fit, as opposed to close-ended questions where the possible alternative answers are already listed. Open-ended questions fit nicely in qualitative research, while close-ended ones are better suited for quantitative research (Breakwell et al., 2000).

### **4.1.2 Interviews**

Interviews let you ask users about their experiences and preferences with the product (Repstad, 1999). It is a formal, structured event where you directly interact with users, asking them to voice their opinions and experiences regarding the product.

The negative part of interviewing is that persons answering questions are not as problem-oriented as they are in an authentic situation. It is hard to conclude that the interviewees give an operative valuation based on a valuation in an interview. It is highly probable that the interviewees give a valuation that fits with their self-images, their opinions about what other people think of them. For this reason, finding an arena where the real valuation comes into being is essential (Repstad, 1999).

When conducting interviews the respondents might answer not only to the current question but also broaden their answers to include other issues. This is something that should not be looked upon as a bad thing. If the respondents are not allowed to continue talking about a matter that they obviously find important, they might become a bit insecure or annoyed, and consequently more quiet or agitated. Another thing that

is good to consider is that a small informal talk at the beginning of the interview could get the respondent to speak more freely (Repstad, 1999).

As mentioned above, spontaneous conversations at the target site of the product are favourable, but not always possible. Therefore, conducting an interview while simulating an actual arena is a viable alternative. Concurrent, contextual interviewing is when you get information by combining observing, listening and asking the user questions in the context of the user's tasks (Repstad, 1999; Hackos & Redish, 1998; Hom, 1996).

As you would expect, the interview questions themselves are an important part of the interview. Besides open-ended and close-ended questions there are a lot of different types of questions, like general and specific questions, or judgmental and hypothetical questions. A judgmental question is used when you want to know the user's attitudes, opinions, motivations and expectations. A typical judgmental question could start with "What do you think about..." When you ask subjective questions, however, you need to be especially careful to ask them in a neutral manner so that you get the user's real opinion and not what the user thinks you want to hear (Hackos & Redish, 1998).

#### **4.1.3 Scenarios**

Scenarios are descriptions of possible interactions with a product. They can be used as a method of gathering information about how users will use the product to deal with their tasks. A scenario can also reveal specific problem areas and errors. It enables users to look at the features and functions offered by a product. In other words, a scenario can be used to figure out whether a certain function in a product is worth constructing before actually constructing it (Faulkner, 2000).

#### **4.1.4 Think aloud protocol**

The best understanding of how users do their work usually comes from talking with the user about the tasks while they are being performed. Even immediately after a task, users may not be able to recall why they did something. It is then good to encourage them to "think aloud". At every step in performing the task the users should explain what they are doing and why. The think-aloud protocol is useful when you want to get the user's reasons and decisions while performing a task. This

solves the problem of finding a suitable arena for an interview, as discussed in Section 3.3.2 (Hackos & Redish, 1998).

### **4.1.5 Paper prototyping**

Hackos & Redish (1998) state that one of the best ways to explore and encapsulate design ideas is through prototypes. A prototype can be an easily changeable sketch or mock-up of at least a part of an interface. These kinds of prototypes are often called low-fidelity prototypes. High-fidelity prototypes require a functioning, almost complete product.

Every kind of prototype has its advantages and disadvantages. Paper prototypes or sketches have several advantages, the main ones being that they cost little to produce and that they are fast and easy to create and to change. The latter lead to the psychological advantage that users might be more prone to change the prototype than if it was a larger and nicer-looking mock-up. The main disadvantages are that they usually only show some of the final functionality and that a human being is required to be the *Wizard of Oz*. The Wizard of Oz is a method where the user is presented with what appears to be a working system, but behind the system (where the user can not see) is a human being. Another disadvantage that Hackos & Redish (1998) mention is that these paper prototypes may lack what is called *face validity*, which means that the user might not take them seriously enough. Prototypes that are low in fidelity may also bias users to rate them lower in usability than a prototype of higher fidelity.

Low-fidelity, non-interactive, computer-drawn prototypes have face validity and will then give a more serious impression. But the disadvantage is the reverse, users might not be inclined to change anything because it already looks nice enough and they might believe that there are several months of research behind these prototypes (Dumas & Redish, 1993).

Creating just one set of prototypes is probably far from enough. To get the design right requires iterative prototyping, redesign after redesign. This is apparently not cost-effective, though Hackos & Redish (1998) recommend at least two or three refinement steps. Nielsen (2000:1) has the same opinion, which is explained below. To keep the cost low, Nielsen (1994:1) proposes constructing a horizontal or a vertical prototype instead of a full system. In a horizontal prototype, the whole interface and main functions are visible and "working", but no sub-

functionality is included. In a vertical prototype, focus is on one or two specific tasks and no other functionality is included. A scenario prototype (Dumas & Redish, 1993) differs from the other two types by being task oriented. It can include some functionality from different parts of the system. Nielsen (1994:1) recommends the use of scenario prototypes as a tool for evaluating usability because they are easily changeable on the basis of user feedback. But there is a problem with these partial prototypes. Since they imitate only part of a product; this may lead the test team to overestimate its usability.

#### **4.1.6 Cognitive walkthrough**

Cognitive walkthrough is a method where scenarios are constructed from an early prototype and then letting the user "walk through" the interface. Usually, the main focus of the cognitive walkthrough is to establish how easy a system is to learn. The focus is on learning through exploration. The walkthrough process involves examining each individual step in the correct action sequence and trying to tell a plausible story about why the user would choose that action. The prerequisites are to act as if the interface was actually built. Bottlenecks, where the interface blocks the user from completing the task indicate that something is missing in the interface, are easily found using cognitive walkthrough. During a walkthrough it is fairly effortless to find complicated paths in the interface. This indicates that the interface needs a new function that simplifies the paths (Faulkner, 2000).

#### **4.1.7 Heuristic evaluation**

Heuristic evaluation involves having a small set of evaluators conduct a systematic inspection of a user interface and judge its compliance with a number of generally accepted usability principles. Heuristic evaluation is performed by having each individual evaluator inspect the interface alone. Only after all evaluations have been completed are the evaluators allowed to communicate. This procedure is important in order to guarantee independent and unbiased evaluations. But the heuristics cannot consider every possible system ever imagined. They are broad-based methods and should be considered as such. However, they are good to use when designing a system and making sure that nothing has been forgotten. Using heuristics will not solve all the usability problems but will highlight some issues and improve usability in the whole. An example of a usability heuristic is Ben Shneiderman's "*Reduce short-term memory load*". This heuristic says that displays should be simple and that it must be remembered that a user has limited storage space for information that can be held

in consciousness at the same time. After utilizing the heuristics to find errors, the evaluators will use a severity rating system to decide which errors that are most important to fix (Faulkner, 2000; Nielsen, 1994:2).

The heuristics used in this study are Nielsen's ten recommended heuristics (Appendix F).

### 4.2 Users

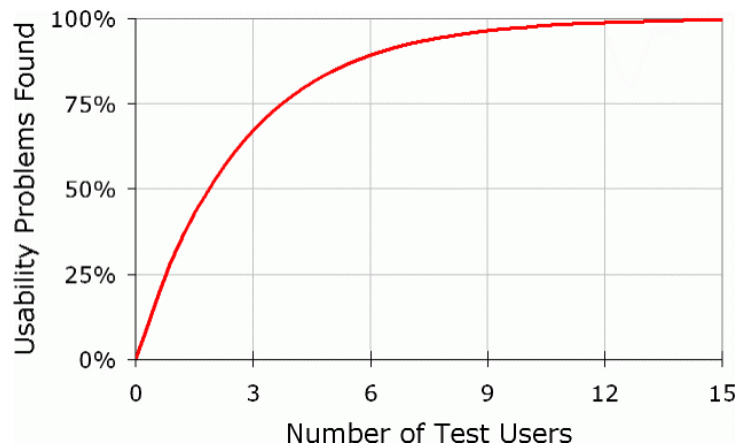
There are two main types of users, primary and secondary users. Only primary users, the ones who perform the task in the application, are discussed in this report. A typical primary user for the tool is a content provider. The secondary users are those who are affected by the tool and the tasks around it, in this case for example a portal user (Hackos & Redish, 1998).

Dumas & Redish (1993, p. 23) state that a "*... test that uses programmers when the product is intended for legal secretaries is not a usability test.*" The users in this study are students, but all of them have a background as amateur content providers or designers of similar applications. Thus they all have something to contribute, some of them have thought about the usability problems when designing an application like this and some of them are used to performing comparable tasks. Dumas & Redish point out that if the test users are more experienced than actual users, a lot of issues might be missed.

But, for every down there is an up. Dumas & Redish (1993) explain that everything is not black or white, for example work experience, general computer experience, experience with similar products (basic and advanced features) are positive characteristics in a user. The participants are supposed to represent real users; hence these characteristics will probably be helpful.

In research studies, researchers often go to elaborate lengths to ensure that they have a random sample of participants from some relevant population. The statistical tests that are applied to the data assume that the participants were selected like this. On the other hand, in common usability testing, you usually have a convenience sample, ie. people from the appropriate population who are available to you (Dumas & Redish, 1993).

Research carried out by Nielsen and Landauer (Nielsen, 2000:1) illustrates why only five users are needed for a usability test (Figure 1).



*Figure 1. Alertbox: Why You Only Need To Test With 5 Users (from Nielsen 2000:1).*

The basic point to notice is that zero users give zero insights. A single user increases the error findings by around 30 percent. Additional user tests will show that there is some overlap in what can be learned, but the increase in insights is almost the same as from zero to one user. With two users you find around 50 percent. A third user raises the insight graph to about 70 percent and as more users are added you will learn less and less. After the fifth user there is no need to continue exploring, there will not be enough to gain from the users. Although the graph tells us that to find 100% of the errors, fifteen users are required, Nielsen suggests that we gain more if a new design is created based on the findings from the five users and tested again with the same five users (Nielsen, 2000:1). The only thing that says otherwise is if one aims for quantifiable research, since it is very hard to do statistical calculations on only 5 users. Dumas & Redish (1993) say that the goal of a usability test is to uncover problems, not to demonstrate the existence of some specific phenomenon.

### 4.3 Testing the goals in practice

The selected line of action is based on Nielsen's discount usability testing (1994:1), a cost-effective method also taking into consideration his opinion about testing with only 5 users (Nielsen, 2000:1).

#### 4.3.1 The participants in this study

It was significant for the study to get users with different skills and backgrounds. All users should be primary users or at least could be regarded as such. One of the easiest groups of people to get access to are university students. They normally have enough time on their hands to participate in studies such as this one. A lot of students are also involved in extracurricular activities and some of them have done work similar to what a typical content provider, a person who writes articles to for instance a portal, outside the university does.

Suitable users would be webmasters for student information sites, preferably at Linköping University (due to its vicinity) because it is easy to get students to participate in research studies. Every user mentioned was contacted by e-mail through that kind of sites. The first contact with the users was made through a kind of pilot test. Four possible users were contacted and asked to fill in a questionnaire. Two of these users promised to participate in the continuation of the study. An additional seven users were asked to take part in the rest of the study (that is, not in the pilot test) and three of them agreed. A total of five users participated in this study. The users are named User1 to User5. The character matrix (see Table 3) displays their characteristics as described in Hackos & Redish (1998). The scales are "*Beginner – Competent – Advanced – Expert*" and "*None – Low – Medium – High*" respectively.

Table 3. Character matrix.

	Sex/Age	Subject matter expertise	Experience of similar applications
User1	M/24	Expert	High
User2	F/22	Competent	Medium
User3	F/23	Competent	Low
User4	M/22	Beginner	None
User5	M/21	Advanced	High

User1 is a former student who works professionally on the development of similar content editors. He is also the main content provider for his student organisation's website. User2 is webmaster and content editor for her own webzine. User3 is webmaster and to a certain extent content editor for her student organisation's website. User4 is technical webmaster for his student organisation's website. User5 is editor-in-chief and content editor for a student magazine.

In the first interview mentioned in section 3.4.4, the users were asked to answer questions (Appendix C) about what experience they have in the topic of interest. The conclusion of this interview was that the users in this study may not be typical content providers but they are close enough for a study of this magnitude. Most of the users have worked with or are working with non-profit student-related content providing, and even though they have not done it on a business level, the tasks are similar.

User1, User2 and User4 are acquainted with the researcher. The user's acquaintance with the researcher could have made them more open and willing to answer questions and think out loud than to a stranger. But this could also have made them want to live up to the researcher's personal view of them and their opinions (Repstad, 1999).

### **4.3.2 Task analysis**

The tool used in this study is intended to be a stand-alone application. There was already a web prototype for the tool. It was made after interviewing employees from a major news and television network company in Stockholm who uses similar products for their web site. The web prototype was used to make a rough sketch over a number of conceivable tasks for the tool. Subsequently, these tasks were merged with the needs that the group currently working on the meta-database had. A requirements document was made from these prerequisites, which was used to make the first drafts of the tool.

### **4.3.3 Pilot test**

The study begins with a kind of pilot test in the form of an email questionnaire. Two participants answered this questionnaire, namely User1 and User5. The REAL goals and measurable objectives in Table 2 were used to determine the questions for the questionnaire.

The test was made to check the validity of the conceivable tasks that were set by the task analysis and to find out if the chosen type of user

was the correct one for the study. The questionnaire is included in Appendix A and contains a scenario, some requests for suggestions in the form of open-ended questions and the conceivable tasks that might be included in the end product functionality. The participants were asked to rate these tasks considering their necessity. The rating was between 1 and 10, where 10 was the most necessary rating. This rating was used to make a first low-fidelity prototype of the product. A print of the first extensive prototype, a low-fidelity paper prototype is included in Appendix E. This prototype was made to gain changeability in the conceivable tasks. A scenario was devised to fit the prototype; it is found in Appendix B.

### **4.3.4 First test suite**

The tests were carried out over a period of two weeks in spring 2001. The first test suite was carried out over two days, three of the user tests on the first day and two on the second. The second test suite was carried out over a three-day period, two on each of the first days and one on the last. This was because the researcher felt that three tests during one day was too much to handle. Tiredness could have an effect on the test results. The tests were performed in the same room each time. The tests took approximately 2 hours each.

In the first test suite, five users were asked to perform a cognitive walkthrough with the researcher based on the scenario and the paper prototype. In return they got a free ticket each to the cinema. The test started with an informal talk to make the users feel a bit more at ease. The researcher explained that the evaluation was of the tool and the technique, and not of the users. The users were informed about the think-aloud protocol and asked if they understood what was required of them. Then, the users were to answer some pre-test questions to learn about their backgrounds and level of experience in the usage of similar tools. Subsequently, the users went through the scenario with the researcher and completed the tasks. During the session, the researcher measured time and approximate error rate, to pinpoint the *learnability* and *efficiency*. The error rates were labelled "None", "Minor", "Major" and "Fatal". The session ended with a follow-up interview containing some questions to pinpoint the *relevance* and the *attitude* (Appendix D). This approach corresponds with the REAL goals and the measurable objectives in Table 2 in chapter 2.3.

The questions have been translated into English in this document by the current author. The relevance questions were the following (Relevance 1 and 2):

- *Do you think that the tool will solve the tasks that you as a web editor would like to perform?*
- *On a scale from 1 to 10, where 10 is the highest rating, how many of the possible tasks that a web editor would like to perform does this tool solve?*

The attitude questions were these (Attitude 1 and 2):

- *Did you like using the tool?*
- *On a scale from 1 to 10, where 10 is the highest rating, how would you rate your overall opinion of the tool?*

Two questions considering the users opinions about the learnability of the tool were asked (Learnability 1 and 2):

- *Did you think that the tool was easy to learn?*
- *How long would you reckon it would take to learn the whole tool (in minutes)?*

One question about efficiency was asked:

- *Do you think that you managed to solve the tasks with a minimum of fuss?*

The practical issues that the users had comments about were taken care of before the next test suite.

#### **4.3.5 Second test suite**

A new task-based scenario (Appendix C) and a new prototype (Appendix E) were made. The prototype was composed of printed screenshots, but otherwise built in the same way as an ordinary paper-prototype. This prototype was made to gain face validity, to make the user feel as if it was the beginning of a real product they were testing. The second test suite began with a cognitive walkthrough of the mentioned scenario in which the users were asked to do their own error rating for each task. Meanwhile, the researcher approximated his opinion of the error rate for each task. This was done to get the same validity on these ratings as in test suite 1 and to make sure that the user's understanding of error ratings was correct.

But the tasks were not the same as in the first test suite since major parts of the tool had been rebuilt. The time factor was not measured due to technical problems. The test concluded with a brief interview, with questions pinpointing the user's *attitude* towards the tool and the *relevance* of the tool (Appendix D). The questions that were asked were

the same as those mentioned above, excluding the last question about efficiency. This question was substituted with the user's own error rating for each task.

### **4.3.6 Heuristic evaluation and third test suite**

A heuristic evaluation of the second prototype was made the day after the last user test, to ensure that nothing was forgotten and to improve the quality. Only User1 evaluated the prototype on this session, due to lack of time. There was supposed to be a third test suite, an evaluation on a high-fidelity prototype or a full-working application, but time posed major constraints.

## 5 Results

In this section the main results of the study are presented. A complete presentation of all results can be found in Appendix G.

### 5.1 Results from the study

While there was consensus amongst the subjects on most of the measurements, there were some discrepancies.

*Table 4.* Answers to interview questions (the first row in each case is test suite one and the second is test suite two).

	User1	User2	User3	User4	User5
Relevance1	Yes	Yes	Yes	Perhaps	No
	Yes	Yes	Yes	Perhaps	Perhaps
Relevance2	8	10	7	8	3
	9	10	8	7	3
Attitude 1	Yes	Yes	Yes	Yes	No
	Yes	Yes	Perhaps	Yes	Yes
Attitude 2	7	10	7	8	4
	7	8	8	8	8
Learnability 1	Yes	Yes	Yes	Yes	Yes
	Yes	Yes	Yes	Yes	Yes
Learnability 2	60 min	60 min	10 min	5 min	90 min
	10 min	15 min	15 min	10 min	90 min

Relevance 1 had to do with the suitability of the tool for the task in hand. Table 4 shows that three out of five subjects answered positively in both test suites while the two others were more reserved. From this we can assume that the users find the relevance of the tool to be high. In Relevance 2 the users were asked to rate the coverage of the tool with respect to common web editing tasks. Here, a tendency was detected showing that most of the users find that the coverage has remained the same or increased between the prototypes.

The same results appear for the users' attitudes towards the tool. In Attitude 1 the users were asked whether they liked using the tool or not. Four out of five subjects answered positively in both test suites. In Attitude 2 the users rated their overall opinion about the tool and from the table we can see that only one user disapproved of the changes made in the second prototype.

It can be seen in the table that all users voted in favour about whether the tool was easy to learn or not (Learnability 1). Additionally, it appears that there is a bad connection between their opinion about learnability and their estimation of learning time (Learnability 2). In the first case we see that all subjects agreed that the tool would be easy to learn. However, when it came to estimating the time it would take to learn the system, the subjects differed considerably, the extremes of the scale being 5 and 90 minutes, respectively. Obviously, the subjects had varying attitudes as to the learning investment required for a tool of this kind.

Another factor that can readily be seen in Table 4 is that User5 differs on almost all accounts from the other subjects. A follow-up question after the interviews showed that this is due to the fact that he is used to completely different work procedures and tasks.

The efficiency measurements, that are missing in Table 4, could not be compared since they differed too greatly in the two test suites. This problem is discussed in section 5.1. In the second test suite, both the researcher and the users carried out the error rating. The result of this was that the users' opinion about their own error rate did not differ from the researcher's opinion (Appendix G).

To sum it up, four of the five users, as mentioned above, liked using the tool since they believed the work paths kept the flow in the tool and were logical. They thought that combining the tasks in the same tool made it different from other such tools that they had worked with. Some users thought that the tool was very practical and at the same time very mechanical, they missed some features that would make the tool more like an environment where the users can follow different work paths and create different sorts of articles.

But, on the other hand, they believed that if you work with this tool everyday, you would need it to be mechanical. One user called it *harmless*. Everybody liked the what-you-see-is-what-you-get-interface, since it gave a hands-on feeling for what you are doing. One user needed other means that were thought to be outside the area, to complete his daily tasks. Another user simply disliked that there were no hidden games in the tool. Everybody thought that the tool would be easy to learn, even for people with low computer skills.

## 6 Discussion

When it comes to the *method*, I found it hard to convert the REAL goals (Löwgren, 1993) to qualitative objectives and find the optimal measuring system. Some feedback was picked up from co-workers and my supervisor, but when looking back, this was sadly not enough. Some changes were made, but several iterations would have been needed to get it right.

The efficiency questions could not be compared since they differed too greatly in the two test suites. If these questions had been the same, a calculable comparison of the efficiency could have been made. If carried out correctly, a difference in error rate could have been found. A test comparing the users' error rate and their time to complete the tasks might have shown whether time and error rate are good ways to measure efficiency. In the second test suite the time was not measured. When carrying out the test with the first user, it became clear that an assistant would have been a great help, but such a person was not available at the time and as a result the time could not be measured. At this instant, measuring was not considered vital. Afterwards, however, it would have been rewarding to compare the measured times with the corresponding error ratings. Consequently, the tasks used in efficiency testing should be more alike in the different test suites, but this could prove difficult. A usability study is always focused on getting results that should shape the product, and if a task flow gets changed, then the efficiency study fails.

Learnability could perhaps have been studied further by measuring the time it took for the users to "completely" learn the tool, but this would have required the whole tool to be complete, since many factors are involved in human learning. But, as several of the users were used to using new computer applications, they should have had enough experience to give a fairly accurate estimation of the learning time of the tool.

The study was conducted using only five users. As mentioned previously, this makes it hard to do statistical analyses and hence get results that have statistical validity. If more users had been available, a within-subject design, having a control group carrying out the tasks, could have been used to make it easier to use statistical tests such as the T-test. But this tactic fails if the prerequisites change, which they did in this study (the tasks changed). Another approach could have

been to use a between-subject design, with two equally large groups carrying out the same test. But even if this would produce a lot of data to work with, it would not provide much specific help in the implementation of the tool. Nielsen (2000:1) states that five users are enough when conducting usability studies, and I agree.

It can also be argued whether using paper prototypes are the proper means for carrying out usability studies. The answer is both yes and no. It doesn't really project the typical computer feeling on the user and what I've measured is really the usability of the paper prototype. In the end, most usability results should be analogous when using a prototype or using a fully featured tool. The paper prototypes used in the study were made in view of tasks instead of a full-blown scenario, which was recommended by several experts (see section 3.3.5). But it is difficult to say whether this has made any impact on the results.

On the other hand, the use of paper prototypes seems to have made people more inclined to change both small details and full tasks in the tool. The interview questions were apparently fully understood and the answers were of the type described above. I believe that the use of a scenario gave a more in-depth understanding of the tool and that the cognitive walk-through provided a lot of room for discussion. Using the efficiency rating made discussion topics simple; several of the test sessions did not end within the assumed time frame, since the users had much to discuss.

Despite this, during the interviews, none of the users showed signs of tiredness or inattentiveness, apart from one user. User2 gave highly positive ratings throughout the first test suite. When interviewing this subject, it appeared that she may have been distracted, since she had some emotional problems at that time and may therefore have tended to be less incisive in her judgment than would normally be the case.

As far as the *results* are concerned, *relevance* was measured by letting the user subjectively rate the tool or to be more precise, the tasks that the users had performed including hints they have picked up about the peripheral tasks of the tool. The result of this is still only a hint as to whether the tool performs the correct tasks to meet the users' needs. Measuring the relevance of different tasks or function in the tool might give better understanding. Also, it might be possible to devise a more objective measuring scheme, like error rating. Relevance has a relatively low level of subjectivity compared to attitude. The users base

their attitude towards the tool on their own feelings while they build their opinion of the relevance of the tool on pre-conceived notions from their surroundings.

My personal belief is that measuring *efficiency* by letting the user rate errors together with the researcher is advantageous, since it became easier to locate usability issues that had to be solved. Using the four-grade scale provided a straightforward way to prioritize the issues. If comparable tasks are measured in the different test suites, one might use the results to ascertain an increase in quality for the management or customer, but this method is still too shallow to use as a research base. To measure efficiency, by error rating or time measurements, is by and large a good way to spot faults and at the same time give people who needs them some numbers to compare.

Measuring *attitude* is fairly the same as measuring relevance with the difference that attitude is far more subjective. It is questionable whether one could gain from measuring attitude. A better approach is to ask the users if they think that the tool is better to use than it was before changes were made. If they say no, then one might want to try to discern why, and this is nothing that can be measured in a quantifiable way.

The research approach used in this study is of the traditional kind. I have tried to measure by the book and it appears that some parts of the measuring do not seem to work. From the *learnability* questions, it can be seen that the qualitative answer "yes" differs quite a lot from the quantitative answers that range from 5 to 90 minutes. From this it doesn't seem that you can measure learnability in this way, or that it is relevant to do so. One can argue that if the users get a chance to learn the tool, and that you measure the time of learning, you'll get good numbers and a high transferability. But, this is not formative evaluation, to execute this the tool has to be complete. However, since the users estimates did not differ completely, one can argue that such estimations can be used successfully in a formative evaluation anyhow, not to get some sort of accurate results, but to provide an early indication of what might be right or wrong.

When comparing the learnability results with the character matrix (Table 3), one can see that the two users (user3 and user4) with least knowledge of similar tools have the lowest estimate of learning time. It is my belief that this is neither coincidence nor is it that they are not

experienced enough in using other computer software. All the users are webmasters or content editors of some kind, with good knowledge of different software. If further testing had been carried out, the real learning time might possibly be in the region of these users' estimates.

This would support the view that when conducting usability evaluations, one should not focus just on the typical users but include other kinds of users as well. The implementation of software should, naturally, always include potential users. I was lucky to get such wide variety of users in this study, especially since User5 had a completely different view of how a content editor should work. This made the design decisions a lot easier, even though the study results look odd; yet another reason not to use closed questions and measurable methods.

Contrary to what I thought at the beginning of the study, the fact that I knew some of the users beforehand seems to only have brought good. These users were prone to tell me everything they felt and thought about the tool, without feeling troubled with me taking notes or asking tricky questions. This might be due to their openness towards people in general, but nonetheless letting them contribute gave results.

As mentioned above, when discussing the trustworthiness of the study, if learning time were to be measured when the tool was finished, one would get good transferability. In regard to what this study has produced, moving the study into another context would probably not give similar results. That makes both the transferability and the dependability of this study quite low. On the other hand, the credibility of the results is still high, since the users had different backgrounds and experience, and still had similar individual results. And even though some of the users were known to me, the confirmability should be quite high, since the known users answered in much the same way as the others.

At the beginning of this report it is said that the intent of the study is to find out whether the REAL usability goals can be turned in to measurable objectives and if they are relevant for a practical usability process. It appears that, by and large, any usability goal can be turned in a good measurable objective. It is my belief that these measurable objectives can be used fruitfully to give a pointer in qualitative research, but they should not be used only to achieve statistical validity. The aim in usability research is and will always be to make usable products, with

the focus on the product, and not on the process. Consequently, when conducting a usability study, use measurable objectives to get an easier way to compare results; do not use them to make nice charts. When trying to make goals quantifiable, you do not really wish to make conclusions over the whole population, you only want to make your product better to use. The focus is simply on different objectives. Using measurable objectives and comparing the results is somewhat dubious. If you look at the results in this study you can say that the users' estimate of learning time has decreased in the second test suite and hence this second prototype is easier to use than the first. You might say that the users prefer the second one, but this never shows that the second prototype is the final answer.

The whole usability method that has been used in this study and that is used around the world proposes a way to measure the usability of a system, but it does not reveal how a system might be changed in order to improve its usability. It will be required to find a measure whose aim is not merely to help in creating nice charts or giving hints on what to do the second time around, but rather gives direction. I suggest that the proper wording will be usability *target* instead of usability *goal*, because the process is about aiming in the right direction. Reaching the goal will be unachievable, but getting as close as possible is already the real goal.

## 6.1 Summary

To summarize the report, the REAL usability goals can be used successfully in a practical usability process and what this study has told us about making REAL into measurable objectives is the following; *Relevance* is hard to measure but it is still important since it gives a good hint about where the usability process is heading. *Efficiency* is easy to measure and lets us spot faults, but it is hard to use in a comparative study. Measuring *Attitude* is, compared to the others, almost futile, and we gain very little when we try to. *Learnability* could be easy to measure, if done properly, and it would probably give us a lot to work with. The conclusion of this is that if Attitude is to be a normal usability goal and Relevance, Efficiency and Learnability are used to create measurable objectives, it would prove beneficial in the usability process.

### **7 Future work**

While the objectives of this study have been achieved, more needs to be done. One of the most crucial issues would be to determine whether a usability method based on both measurable objectives and qualitative goals in an optimal combination can be devised. This would give a usability study an extra dimension, which may possibly produce more and/or better results.

## 8 References

Breakwell, Hammond & Fife-Schaw. (2000). *Research Methods in Psychology*. London: SAGE Publications Ltd.

Dumas, J.S. & Redish, J.C. (1993). *A Practical Guide to Usability Testing*. Exeter, UK: Intellect Books.

Eason, K. D. (1984). Towards the Experimental Study of Usability. *Behaviour and Information Technology*, 2(3).

Faulkner, X. (2000). *Usability Engineering*. Basingstoke, UK: Houndmills.

Hackos, J.T. & Redish, J.C. (1998). *User and Task Analysis for Interface Design*. New York: John Wiley & Sons.

Hom, J.T. (1996). *The Usability Methods Toolbox*.  
<http://jthom.best.vwh.net/usability/> (Last checked 020705)

Kindborg, M (1999). Interaktionsmodeller. (in Swedish),  
<http://www.ida.liu.se/~mikki/mdi/interaktion.html>  
(Last checked 020705)

Kirakowski, J. (2000). *Questionnaires in Usability Engineering*.  
<http://www.ucc.ie/hfrg/resources/qfaq1.html> (Last checked 020705)

Lewis, C. & Norman, D. (1986). *Designing for Error* in Norman D & Draper S (1986) *User Centred System Design*. Hillsdale, New Jersey: LEA.

Löwgren, J. (1993). *Human-computer Interaction*. Lund: Studentlitteratur.

Nielsen, J. (1994:1). *Discount Usability Engineering*.  
[http://www.useit.com/papers/guerrilla\\_hci.html](http://www.useit.com/papers/guerrilla_hci.html) (last checked 020705)

Nielsen, J. (1994:2). *How to Conduct a Heuristic Evaluation*.  
[http://www.useit.com/papers/heuristic/heuristic\\_evaluation.html](http://www.useit.com/papers/heuristic/heuristic_evaluation.html)  
(Last checked 020705)

Nielsen, J. (2000:1). *Alertbox: Why You Only Need to Test With 5 Users*. <http://www.useit.com/alertbox/20000319.html>  
(Last checked 020705)

Nielsen, J. (2000:2). *Usability Metrics: How good are you*.  
<http://www.zdnet.com/devhead/stories/articles/0,4413,2321306,00.html>  
(Last checked 020214)

Nielsen, J. (2001). *Alertbox: Usability Metrics*.  
<http://www.useit.com/alertbox/20010121.html> (Last checked 020705)

Repstad, P. (1999). Närhet och distans (Kvalitativa metoder i samhällsvetenskap), (in Swedish), Lund: Studentlitteratur.

Trochim, W.M.K. (2001). *Qualitative validity*.  
<http://trochim.human.cornell.edu/kb/qualval.htm> (Last checked 020705)

## Appendix A: Pilot test questionnaire

Hej!

Jag heter Martin Karlsson. Jag har läst på ett antal olika program på LiU och har nu bestämt mig för vilken examen jag ska ta, därmed är jag nu mitt inne i ett examensarbete på företaget Xpedio AB. Examensarbetet går ut på att ta fram ett verktyg för webbredaktörer och det är där ni kommer in i bilden. Du som får detta mail har förhoppningsvis varit redaktör för en tidning eller en webbsajt. Om nu känner att ni har några minuter över så får ni gärna svara på nedanstående korta frågor och skicka dem till mig innan torsdag. (Svaren behandlas anonymt.)

Jag håller just nu på att utforma vilka krav som finns för ett redaktörsverktyg och skulle vilja ha er hjälp att verifiera alternativt motsäga dessa krav.

Uppgift:

Du är webbredaktör för en webbportal. Du har i uppgift att lägga in artiklar på webben (exempel kan ses på exempelvis <http://www.spray.se/nyheter/> ). Det är ganska mycket information som ska läggas in så du behöver troligen hjälp både av ett verktyg och av andra personer. Sätt viktighetsgrad (1-10) på följande delar av ett tänkbart verktyg som ska hjälpa dig i ditt arbete (flera delar får ha samma viktighetsgrad):

- Skapa artikel (rubrik, artikeltext, bild, länk till ytterligare information)
- Ändra innehållet i artikeln
- Kategorisera artikeln
- Ta bort artikel
- Skapa kategori (Inrikes, Utrikes, Sport, etc.)
- Ändra kategori
- Ta bort kategori

- Skapa roll (låt andra personer få rättigheter att bli redaktör)
- Ändra roll, Ta bort roll

### Frågor:

1. Anser du att det saknas viktiga delar i det tänkta redaktörsverktyget och i så fall vad?
2. Finns det delar som du finner onödiga och i så fall vilken/vilka?
3. Har du använt ett webbredaktörsverktyg tidigare och i så fall vilket/vilka?
4. Kan du tänka dig att utvärdera det här verktyget under utvecklingsprocessen samt när det är färdigt?

## Appendix B: Test suite 1

### Minnesanteckningar

Berätta om Xpedio och device-independence, portalen, I-mode, WAP, etc. Berätta vad det är som ska testas, att det är ett webbredaktörsverktyg som ska köras på i Windows på en PC. Testet är uppdelat i två delar ...

Det är verktyget som utvärderas och inte du som användare. Du får avbryta testet när du vill om du känner dig illa till mods. Fråga om det är något som du inte förstår eller om du är nyfiken. Under försöket får du mer än gärna berätta vad du tänker, då detta hjälper oerhört i utvecklingsprocessen. Då försöksmaterialet är på engelska hoppas jag att du berättar om du kör fast på grund av någon språklig miss.

### Pre-test questions

1. Vilka erfarenheter har du av webbpublicering?
2. Har du använt ett webbredaktörsverktyg tidigare?
3. Har du några frågor innan vi börjar?

### Scenario

Du jobbar som webbredaktör på CNN Europe. De använder sig av Xpedios portalsystem för att publicera nyheter på webben och andra plattformar.

Du har ansvaret över CNNs vädersidor och det har kommit in en nyhet under natten. Skapa en ny intern artikel under Weather-kategorin och lägg in nyheten under den. Lämpliga keywords är "weather", "animals" och "seals". Du väljer en mall (template) som passar för vädernyheter. Artikeln ska publiceras för alla plattformar. Kopiera texten till dem. Publicera inte nyheten nu, utan spara den opublicerad tills vidare.

Din chef kommer in och ger dig en webbadress (en URL) i handen. "Publicera den här nu!", ropar han och stressar vidare. URLen ska publiceras under Weather-kategorin. Skapa en ny extern artikel och publicera den mellan den 16:e och den 17:e maj.

Tiden är nu inne för att släppa dagens nyhet på huvudsidan. Du väljer artikeln du sparade tidigare och publicerar den.

Chefen stormar in på kontoret och gormar att den externa artikeln (URLen) som han gav order om att publicera tidigare ska tas bort och absolut inte publiceras igen, då den inte innehöll något väder-relaterat. Du blir förvånad men tar givetvis bort artikeln.

### **Follow-up interview**

1. Tror du att det här verktyget löser de uppgifter som du som webbredaktör vill utföra?
2. På en skala från 1 till 10, där 10 är högst, hur stor del av de uppgifter som en webbredaktör kan tänkas ha löser detta verktyg, tror du?
3. Tyckte du att du lyckades lösa uppgifterna utan problem?
4. Tyckte du om att använda verktyget?
5. Vilket är ditt allmänna omdöme av verktyget på en skala från 1 till 10, där 10 är högst?
6. Tyckte du att verktyget var enkelt att lära sig?
7. Hur lång tid tror du att det tar att lära sig verktyget?

## Appendix C: Test suite 2

### Tasks

Du jobbar som webbredaktör på CNN Europe. De använder sig av Xpedios portalsystem för att publicera nyheter på webben och andra plattformar.

Du får ett antal uppgifter att lösa. Efter varje uppgift stannar du upp och ger ett betyg från 1-4 (där 4 är högst) på hur stort problem du tyckte att det var att lösa uppgiften.

1. Skapa ny artikel under Weather/World.
2. Fyll i rubrik, ingress och artikeltext.
3. Byt redigeringsläge till Imode.
4. Minska ner ingress och artikeltext, se till att de är låsta.
5. Byt tillbaka till standardläget.
6. Fyll i nyckelorden Weather, Russia, Putin, Floods.
7. Fyll i en lämplig mall för denna nyhet (template).
8. Lägg till en länk till artikeln, fyll i källa och länkade ord.
9. Spara artikeln opublicerad.
10. Skapa en ny länk under samma kategori som ovan.
11. Fyll i rubrik, källa och URL.
12. Fyll i nyckelorden Weather, Russia, Seals.
13. Publicera länken mellan kl 15.00 den 25 maj och kl 15.00 den 10 juni.
14. Publicera artikeln du skrev tidigare.
15. Ta bort länken du nyss publicerade.

### **Follow-up interview**

1. Tror du att det här verktyget löser de uppgifter som du som webbredaktör vill utföra?
2. På en skala från 1 till 10, där 10 är högst, hur stor del av de uppgifter som en webbredaktör kan tänkas ha löser detta verktyg, tror du?
3. Tyckte du att du lyckades lösa uppgifterna utan problem?
4. Tyckte du om att använda verktyget?
5. Vilket är ditt allmänna omdöme av verktyget på en skala från 1 till 10, där 10 är högst?
6. Tyckte du att verktyget var enkelt att lära sig?
7. Hur lång tid tror du att det tar att lära sig verktyget?

## Appendix D: Interview questions table

Table 5. Types of interview questions.

Type	Definition
Closed	Structured, limits responses.
Open	Broad, no fixed responses. Invites discussion.
General	Focuses on the big picture
Specific	Focuses on details
Judgmental	Asks for opinion
Hypothetical	Speculative

Table 6. Interview questions in this study.

Question	Type
Do you think that the tool will solve the tasks that you as a web editor would like to perform?	Closed, General, Hypothetical
On a scale from 1 to 10, where 10 is the highest rating, how many of the possible tasks that a web editor would like to perform does this tool solve?	Closed, Hypothetical
Did you like using the tool?	Open, Judgmental
On a scale from 1 to 10, where 10 is the highest rating, how would you rate your overall opinion of the tool?	Closed, Hypothetical
Did you think that the tool was easy to learn?	Open, Judgmental
How long would you reckon it would take to learn the whole tool (in minutes)?	Closed, Speculative
Do you think that you managed to solve the tasks with a minimum of fuss?	Closed, (Hypothetical), Judgmental

This page intentionally left blank.

## Appendix E: Low-fidelity prototypes

Two example scans of the first low-fidelity prototype is shown below. The functionality that is shown in these scan are then shown in two screenshots of the second low-fidelity prototype.

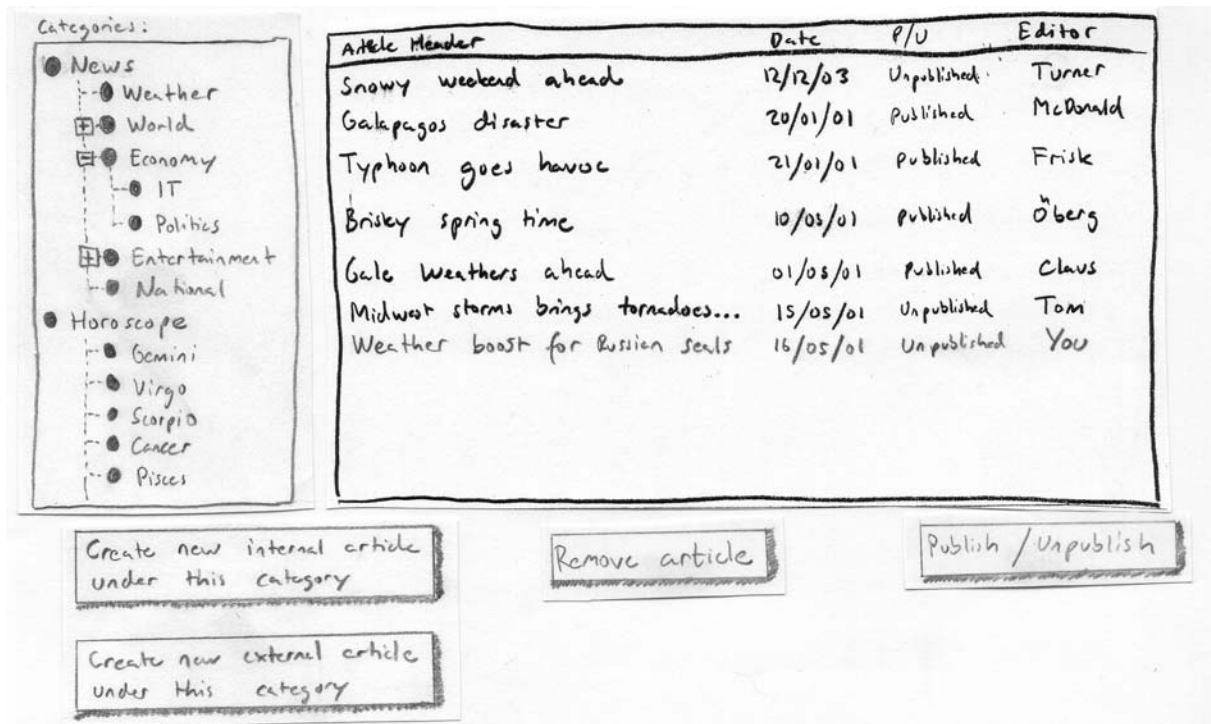


Figure 2. Main window with category and article lists in prototype 1.

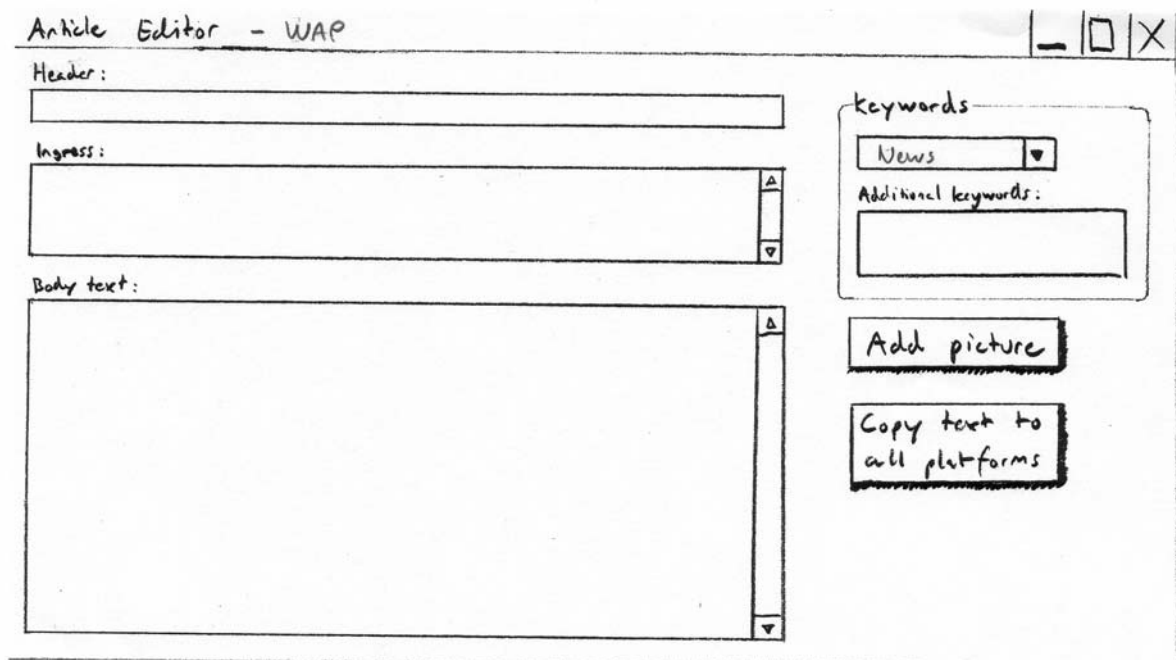


Figure 3. Article editor window in prototype 1.

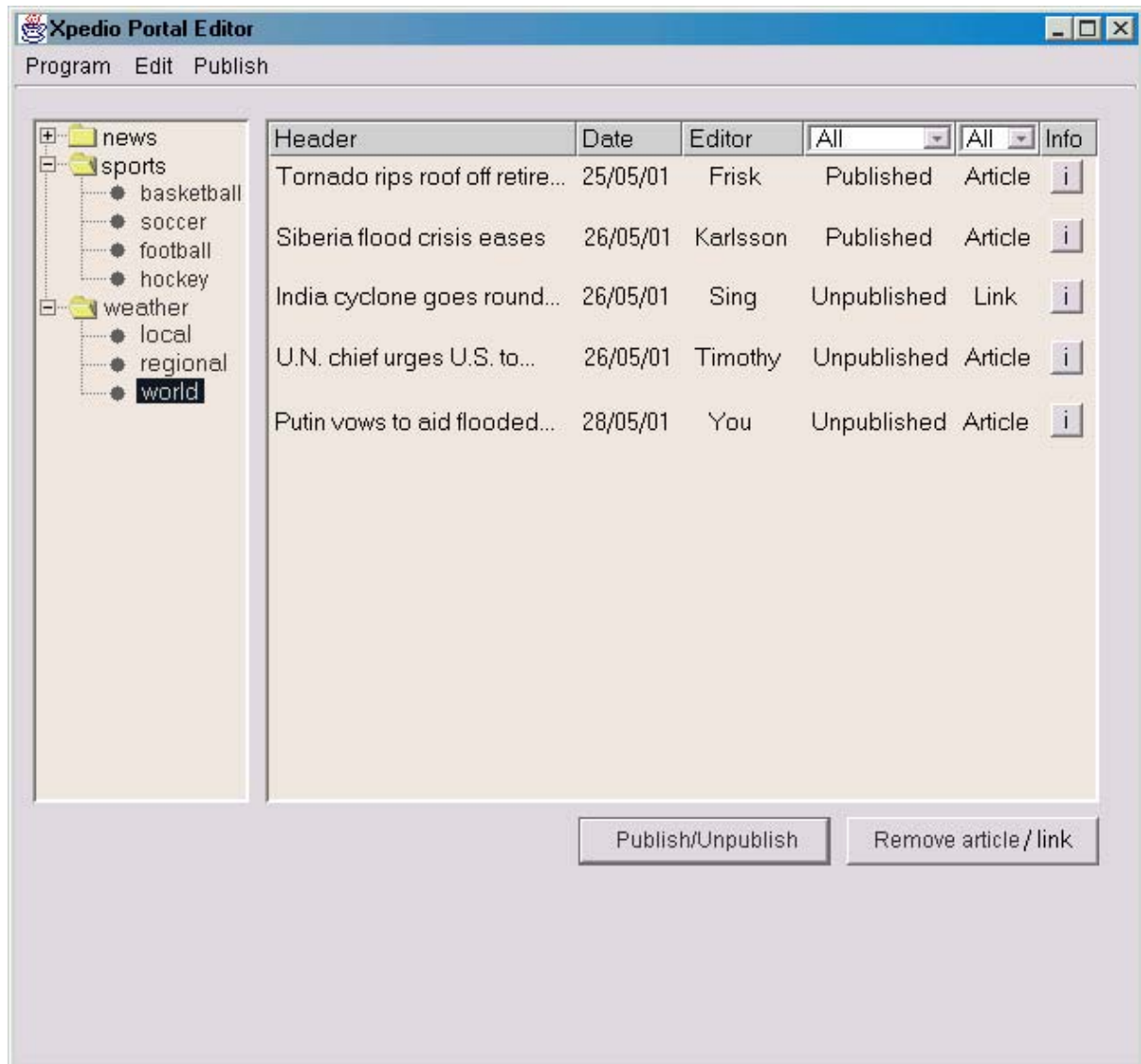


Figure 4. Main window with category and article lists in prototype 2.



Figure 5. Program menu for prototype 2.

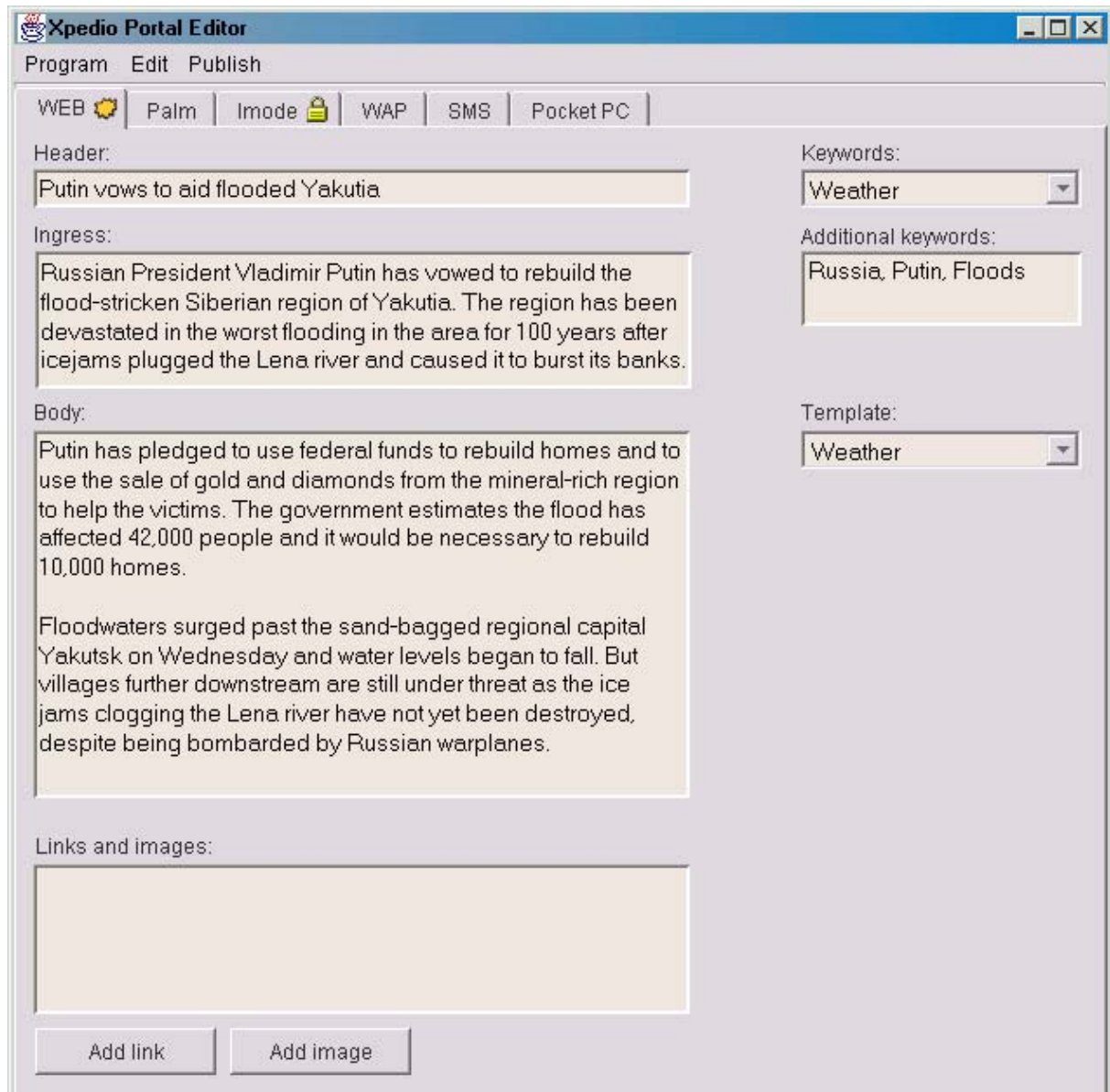


Figure 6. Article editor window in prototype 2.

## **Appendix F: Nielsen's usability heuristics**

### **Visibility of system status**

The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.

### **Match between system and the real world**

The system should speak the users' language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.

### **User control and freedom**

Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.

### **Consistency and standards**

Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.

### **Error prevention**

Even better than good error messages is a careful design, which prevents a problem from occurring in the first place.

### **Recognition rather than recall**

Make objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.

### **Flexibility and efficiency of use**

Accelerators -- unseen by the novice user -- may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.

### **Aesthetic and minimalist design**

Dialogues should not contain information, which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.

### **Help users recognize, diagnose, and recover from errors**

Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.

### **Help and documentation**

Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large.

[http://www.useit.com/papers/heuristic/heuristic\\_list.html](http://www.useit.com/papers/heuristic/heuristic_list.html)

## Appendix G: Test result summary

Table 4. Answers to interview questions.

	User1	User2	User3	User4	User5
Relevance1	Yes Yes	Yes Yes	Yes Yes	Perhaps Perhaps	No Perhaps
Relevance2	8 9	10 10	7 8	8 7	3 3
Attitude 1	Yes Yes	Yes Yes	Yes Perhaps	Yes Yes	No Yes
Attitude 2	7 7	10 8	7 8	8 8	4 8
Learnability 1	Yes Yes	Yes Yes	Yes Yes	Yes Yes	Yes Yes
Learnability 2	60 min 10 min	60 min 15 min	10 min 15 min	5 min 10 min	90 min 90 min

Table 7. Efficiency rating in the second test suite.

Task	User1		User2		User3		User4		User5	
1	3	4	1	2	2	1	1	2	2	3
2	1	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1	1
4	3	4	1	1	3	2	1	2	2	1
5	2	2	1	1	1	1	1	1	1	1
6	2	1	1	1	3	2	1	1	1	2
7	1	1	1	1	1	1	1	1	1	1
8	2	1	2	2	3	3	1	2	2	2
9	2	2	2	3	2	2	2	3	1	1
10	1	1	1	2	1	2	1	1	1	1
11	1	1	1	1	1	1	1	1	1	1
12	1	1	1	1	1	1	1	1	1	1
13	1	2	1	2	1	2	1	2	1	1
14	1	1	1	1	1	1	1	1	1	2
15	1	1	1	1	1	1	1	1	1	1

The values correspond to the users rating of the difficulty of each task. Rating 1 means no difficulty whilst rating 4 means great difficulty. The second value in each user column is the researchers approximation of his opinion of the error rate for each task.

The heuristic evaluation gave few negative and many positive results, which could mean that the prototype had evolved enough to be implemented.

*Table 8.* Heuristic evaluation results.

<b>Visibility of system status</b>	A status window is missing
<b>Match between system and real world</b>	Program, Linked Words are bad. The File menu isn't used for filing.
<b>User control and freedom</b>	If an article is removed, it is forever gone. You cannot move articles between categories.
<b>Consistency and standards</b>	Is the Publish Now-button the same as the Publish/Unpublish-switch?
<b>Error prevention</b>	-
<b>Recognition rather than recall</b>	The create-button is misplaced
<b>Flexibility and efficiency of use</b>	It is nice to use an accelerator instead of clicking on the create-button. Different ways to reach the Publish-mode is good.
<b>Aesthetic and minimalistic design</b>	-
<b>Help users recognize and recover from errors</b>	-
<b>Help and documentation</b>	Totally missing.



**LINKÖPINGS UNIVERSITET**

**Avdelning, institution**  
Division, department

Institutionen för datavetenskap

Department of Computer  
and Information Science

**Datum**  
Date

2002-10-17

**Språk**  
Language

- Svenska/Swedish  
 Engelska/English

\_\_\_\_\_

**Rapporttyp**  
Report category

- Licentiatavhandling  
 Examensarbete  
 C-uppsats  
 D-uppsats  
 Övrig rapport

\_\_\_\_\_

**ISBN** —

\_\_\_\_\_

**ISRN** —

\_\_\_\_\_

**Serietitel och serienummer**  
Title of series, numbering

**ISSN** —

\_\_\_\_\_

LiTH-IDA-Ex-Ing-02/25

**URL för elektronisk version**

**Titel**  
Title

Turning Usability Goals into Measurable Objectives

**Författare**  
Author

Martin Karlsson

**Sammanfattning**  
Abstract

Usability is typically measured relative to a user's performance on a given set of tasks. A number of different goals are used as a foundation for usability testing. The aim of this study was to discover whether the usability goals could be turned into measurable objectives.

The most common measures used in usability testing are time, error rate and the user's subjective satisfaction. In this study, a conversion from usability goals to objectives using these measures was performed. User tests were carried out. The test suites were composed of interviews and cognitive walk-throughs of prototypes. The study was conducted with five users. While this is a small number for performing a quantitative statistical study, it has been shown that useful results can be achieved with a small number of subjects.

The results of the study show that usability goals can be turned into measurable objectives. These objectives can be used to give a pointer in qualitative research, but they should not be used only to achieve statistical validity.

**Nyckelord**  
Keywords

Usability, usability goals, measurable objectives, usability testing, REAL, relevance, efficiency, attitude, learnability