

A SURVEY OF EMPIRICAL USABILITY EVALUATION METHODS

GSLIS Independent Study

Peishan Tsai

INTRODUCTION

Usability evaluation methods (UEMs) are methods used to evaluate the usability of a product design, and identify the problem areas. UEM has been a much discussed topic in the human-computer interaction and computer science fields for more than two decades, and a vast amount of research has been and is still being done on the topic. While laboratory based usability testing became the primary UEM in the 1980s; a series of alternative UEMs were developed in the 1990s in the hopes to lower the cost and time required for a formal laboratory testing. Among them, empirical user testing and usability inspection methods have emerged to be the principal types of evaluation methods (Butler, 1996). This paper aims to give the reader an overview of these major empirical UEMs used in practice.

The following sections will first look into the give an introduction to usability and usability evaluation methods; than provide an overview of the most commonly used empirical UEMs in the field: user study methods (usability testing, surveys, etc), and system inspection methods (heuristic evaluation, cognitive walkthroughs, etc). The next section would introduce the evaluator effect that is shared among, followed by the conclusion to this paper.

WHAT IS USABILITY AND USABILITY EVALUATION METHODS?

The primary goal of usability is to have products developed to maximize the users' ease of use. International Standards Organization in the *ISO 9241-11 (1998) Guidance of Usability* defined usability as “[t]he extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.” Jakob Nielsen, in his online column of August 2003, further defined usability by five quality components:

1. Learnability: How easy is it for a user to complete a basic task at their first use of a system?
2. Efficiency: How quickly can a user familiar with the system perform tasks?
3. Memorability: How easy is it for a returned user to reestablish proficiency regarding the system?
4. Errors: How many errors does a user make using the system? How severe are the mistakes, and how difficult or easy is it to recover from the mistakes?
5. Satisfaction: How satisfactory is it to use the product

Nielsen's components defined usability in a broad sense, covering not only the ease of use and the effectiveness of the product, also the subjective user preferences - satisfaction. However, one must remember that satisfaction and likeability does not equate to usability, or even usefulness, of a product. Dumas and Redish (1993, p.4) offers a shorter definition: usability is determined by how “... the *people who use the product* can do so *quickly and easily* to accomplish *their own tasks*.”

A product with usability problems may be difficult, or even harmful, to operate. A system interface or a website with usability problems may lower user productivity, or drive users away. To improve usability, the product developers must understand the users' needs; the context in which they use the product; and how they interact with it. This is where usability evaluation methods come into play.

Nelsen (1995) categorized UEMs into four basic groups:

1. Automatically – in which usability measures are computed by running a user interface specification through special software.
2. Empirically – in which usability is assessed by testing the interface with real users or experts.
3. Formally – in which usability measures are calculated by exact models and formulas.
4. Informally – in which usability measures are obtained based on rules of thumb and the general skill and experience of the evaluators.

Of the four groups, empirical methods are the most often used.

Empirical evaluation methods can be grouped into user study and system inspection methods. This paper will examine the user study methods of: surveys, focus groups, usability testing (Melkus, 1985), and contextual inquiry (Beyer and Holtzblatt, 1995); and system inspection methods including: expert review, heuristic evaluation (Nielsen and Molich, 1990; Nielsen, 1992; Nielsen, 1994), cognitive walkthroughs (Lewis, Polson, Wharton, and Rieman, 1990; Wharton, Bradford, Jeffries, and Franzke, 1992).

USER STUDIES

How would one learn about the ease of use that a user experiences with a product? The most direct answer would be: Ask them. User studies follow the spirit of “ask the users” by studying how the users use a product, and find out what they think about it. There are many ways to collect user response; one can inquire about or observe a user's interaction with a certain system.

SURVEYS

Survey is a popular method to send out inquiries and collect data from a large population in a short period of time. It could be used as a stand-alone UEM or with other methods to obtain additional user information. Surveys come in various length, detail, and format. They could be done over the telephone, in person, over the mail or email (Kuniavsky, 2003), and generate quantitative information.

Over the years, there has been a variety of surveys available for different kinds of information collecting; surveys could be used to collect participant's attitudes toward individual products, or to measure a selected aspect of usability (Dumas, 2003). The design of a good survey requires skill and time. The questions need to be correlated to

what the evaluators want to find out; able to provide reliable results; and have certain validity to the study (Dumas, 2003, 1094).

One important thing to note is that what surveys truly measure is user preferences, not product usability (Nielsen, 1999). In essence, survey questions explore how the users feel about a product – were the instructions easy to read, were the system easy to operate, or would they use the product again; not how they really performed on it. Often times, the questions fall short on delving into the reasons behind the answers given.

Another problem with surveys is that it is difficult to interpret the results. When a user rates an answer, what does it really reflect? Would one user's scale or rating standards equal to another's? Did the participants answer the questions according to how they truly feel, or by what they think the evaluators wish to see? What does it mean when participants say they are satisfied with a product? Does it mean they were able to efficiently complete a task, or that though the system failed them, they were not annoyed by it? Are these measures really quantifiable? What does the resulted statistics mean?

Despite its low effectiveness rating, organizations still identify survey as one of the most widely used methods because of its efficiency in reaching a large sample size quickly (Rosenbaum, Rohn, and Humburg, 2000).

FOCUS GROUP

Focus group (Rosenbaum et al., 2002) originated as a market research method, and has evolved into a technique also used to study human-computer interaction and human factors.

A traditional focus group is done by inviting a small group of end users in to talk about a product, a marketing campaign, or a new product concept. The discussion is presided over by an experienced moderator, and held in a room with a one-way observation mirror. The moderator takes notes of the happenings, leads the conversation into interesting tangents, encourages comments, prevents the discussion to be dominated by few of the participants, and all the while avoid having any affects on the session's outcome.

Some practitioners believe that with well planning, proper guidelines and a good moderator, focus groups can gather valuable usability data. They believe that though it is not suited for comparative, competitive, or bench-marking studies, focus groups can be used to generate ideas, capture and validate user roles as well as tasks and workflows, and validate high level strategy. However, there are also some major drawbacks that led many practitioners to question its validity in gathering useful user data (Rosenbaum et al., 2002).

Rauch (Rosenbaum et al., 2002, 703) stated that "... the quality of the data obtained from usability focus groups is only as good as the quality of the participant selection and the questions asked." The effectiveness of a focus group relies heavily on the participants to share their opinions and the discussions they had. Though the group setting may be more relaxed to the participants, it might lead to groupthink, where creativity and individual ideas are suppressed and members conform their opinions to what appears to be the consensus of the team. The conversation may lose its function to reflect general opinions by becoming biased, or self-censored by participants. And as hard as the moderators may try, they may inevitably influence the discussion; it could be from a subtle tone of the

voice, or from the words they chose to say. There have also been incidents when participants seek approval from the moderators and try to give comments that would please them. Further more, the method only collects user thoughts; it does not offer an opportunity to see the user at work, or learn more about the context of work.

With the shortcomings of focus group in mind, practitioners have altered the method to improve its performance.

Rosenbaum (Rosenbaum et al., 2002) blends focus group with co-discovery method (discussed in the following section) by breaking participants into groups of 2 to 3 after an initial conversation, let them explore the products, and then bring them back to discuss their experience. Lotus Development mixed field observation and focus group methods, and formulated usability roundtable (Butler, 1996). The company invites users to bring in artifacts from their work, sit with product team members around a conference table, and use the samples (sample data or application files) to explain the context of their work. Through the session, the product designers get to see and discuss the work that users do, the major issues they face, and how technology is used in real life scenarios. Focus troupe (Sato and Salvador, 1999) is a technique whereby dramatic vignettes, performed by live actors, are presented to an audience of potential users. The product concept is featured as a prop or as a dramatic element in the act. With the understanding of the implications, applications, and expectations of the product established through theatrical experience, the audience then takes part in structured conversation about the product concept. The premise of the technique is that live theater can create strong shared contexts among participants, and provides a space for them to expand and invest in the prescribed situations. The effectiveness of the method is still to be tested.

USABILITY TESTING

Usability testing is also referred to as laboratory usability testing or empirical usability testing. It is designed for evaluators to observe and record end users interaction with a product by asking a user to perform a task, thinking aloud, using a test system. Usability testing is rated “highest as an effective usability methodology to create greater strategic impact,” and is widely used in organizations (Rosenbaum, Rohn, and Humburg, 2000, 343).

LABORATORY SETTING

The usability lab is, in general, a room fitted with a one way observation mirror, multiple video cameras, and a product to be tested. The one way mirror allows the developers and evaluators to observe, firsthand, a user’s interaction with a system without being in the same room, therefore eliminating the user’s pressure of being observed. The cameras record the sessions, capturing the participant’s actions, facial expressions, remarks, and key strokes. Some usability testing system may also be able to log tester’s mouse clicks and on screen activities (Melkus, 1985).

PARTICIPANTS

The participants are ideally recruited from the target end user groups. Dumas (2003) pointed out that the key to finding people reflecting the target end user groups is to identify and develop a user profile categorizing shared characteristics of current users, as

well as those that would influence a user's product choices. With the user profile, the evaluators can then create a recruiting screener to select the participants.

A well discussed issue in usability testing is the number of participants needed to uncover most of the usability problems. Dumas (2003) cited research studies by Virzi (1990) that show 80% of the problems are uncovered with around 5 participants and 90% with 10; each additional participant after that contributes with fewer and fewer new discoveries. Nelsen (1993) points out that though it might take 15 participants to fully discover all usability problems, 3 to 5 participants would uncover from 70% to 80% of the problems.

However, Lewis (1994) discovered that the probability of problem detection greatly affects the number of problems a tester is able to find. If a product has good usability to begin with – in other words its usability problems are not as easy to detect – there would be a lower average probability of problem detection, thus requiring a larger sample size to discover the majority of problems. He provided measures to calculate the expected proportion of detected unique problems according to the sample size, and the return on investment. He further suggested evaluators to examine data from other usability studies to determine the probability of problem detection and plan for future ones.

TASK SELECTIONS

One of the essential requirements of a usability test is that the tasks selected must be something the users would want to do with the system in a real life scenario (Dumas, 2003). The tasks should include basic tasks that users perform frequently and that would tap into the core functionality of the product; tasks that would probe possible problem areas of the product; task that would explore the components of a design; and tasks that may be new to the users, or may interfere with previous use patterns. The tasks are presented in scenarios that add context to the tasks, and help the participants feel as if they were doing an assignment in a real life situation.

THINK ALOUD

Think aloud protocols might be the most important part, and the most widely used method in usability testing to gain insight into the participant's thought process. It was listed as one of the six major characteristics of a valid usability test by Dumas (2003, 1097). Before the task, the facilitator would instruct the participant on how to think aloud, give a brief example and a chance to practice the protocol. During the task, the facilitator would observe and record the session, and give committal prompts to remind the participants to keep reporting on their thoughts.

Though think aloud may be traced to the thinking aloud protocols in cognitive psychology research, its execution has varied from its theoretical root. Dumas (2003, 1100) pointed out that in cognitive psychology studies, thinking aloud is used to study what is in participants' short term memories, and discourages from reporting interpretations of the situation, or expressions on emotions or expectations. But in usability study, the focus is on users' interactions with the object being tested, and encourages reporting on thoughts, expectations, feelings, and anything else the participants' have on their minds. Borne and Ramey (2000) reviewed and compared the think aloud protocol theory applied in science research and practiced in usability testing. They noted that one of the biggest differences lies in the subject of the study. Cognitive

psychologist uses think aloud to study the human cognition; the tool with which the task is performed is only an apparatus to facilitate the study. In usability studies, it is the system itself being studied. They proposed speech communication as a theoretical alternative.

THE GOOD AND THE BAD

As with all methods, there are several advantages and disadvantages to usability testing method (Simeral and Branaghan, 1997; Dicks, 2002; Melkus, 1985; Rosenbaum, Rohn, and Humburg, 2000; Karat, Campbell, and Fiegel, 1992). People favor usability testing because of how it:

1. Allows multiple observers at one session.
2. Provides a neutral observation point for the product designers and engineers to see firsthand users' interactions with the system, which often has big impact to the developers, and directly affects future product design.
3. Is conducted in a controlled environment.
4. Employs real users to test the system. How people use an application often exceeds the imagination of experts; some problems can only be discovered by real users.
5. Collects and records a wealth of data with the think aloud method, video cameras, and direct observations.
6. Has shown to be able to expose more severe, more recurring, and more global problems.
7. Generates results that may be immediately implemented and focuses on specific changes to improve ease of use or effectiveness of the product.

But there are also several limitations:

1. Although the testing is designed to mirror users' real work situations, it is nonetheless done in artificial conditions; the task may not reflect what the users would normally use the product for.
2. The participants are rarely a full representation of the target population.
3. The test results do not prove that a product works or that a product is useful; only that the assigned tasks can be completed.
4. Usability testing is inadequate for detecting problems that does not harm product performance, but may affect user's perception of product quality.
5. The test focuses on the chosen tasks, which may limit the evaluation.
6. The laboratory testing requires experimental controls and someone with expertise on the procedure; the facilitator needs to be experienced with the system and the think aloud protocol to smooth the progress along.
7. Usability test is costly. The laboratory is a hefty investment to build and maintain, easily making usability testing the most expensive option among UEMs.
8. The think aloud process is unnatural, distracting, and strenuous to the participants.

ALTERNATIVE METHODS

Of all the limitations of usability testing, cost seems to be the biggest issue that concerns practitioners. Many researchers and authors have suggested simplified methods to do usability testing with smaller budget. Krug (2000) proposes cutting cost by doing the test in any conference room or office with fewer participants, and, instead of hiring a usability expert, have someone who is familiar with the usability test procedures to conduct the test. Nielsen (1994) also suggests forgoing the video recording to further save costs.

Rowley (1994) tackles the issue of usability labs high overhead and maintenance costs with the idea of mobile usability testing; bring the usability test to where the users are, and conduct it in user's environment. This method also allows a development team to collect valuable data from a widely distributed customer base. Another advantage to this method is that the testing will be done in an environment familiar to the users, thus providing context that is closer to real life scenarios.

Co-discovery (or co-participation) (Wilson, 1998) is a method where two participants work collaboratively, and thinking aloud, on the tasks in a usability testing session. Compared to traditional usability testing, co-discovery makes it much more natural to vocalize one's thought and behavior by giving the participant a partner to discuss and converse with. This is especially true when testing an application where people work together. The method also appears to be faster and easier to conduct. But the additional participant in each session means the need for more participants and logistic arrangements for the decided number of testing sessions. There is also a need for more careful screening and pairing of the participants, and the evaluators must be aware of the different learning, verbal, and cultural styles can affect the generated results.

CONTEXTUAL INQUIRY

Raven and Flanders (1996, 2) defines contextual inquiry as “a qualitative data-gathering and data-analysis methodology adapted from the fields of psychology, anthropology, and sociology.” It is a field research method wherein usability evaluators go to the users' workplaces, observes them at work, and asks questions regarding to the work content, process, or product usage. Several evaluators may observe different users at the same time. The data is gathered, compared and shared among product development team members after the observation.

Contextual inquiry is different from an interview because instead of a question and answer session, the data gatherer and the user form a partnership to explore the issues together. Beyer and Holtzblatt (1995) use “apprenticing” to describe the partnership between the product designer and the customer; the product designer follows the user to observe and learn from the user's work context. The process may take hours to months to complete, depending on the user's work life cycle. Raven and Flanders (1996) also stress the importance of basing the contextual inquiry on a focus, a perspective or a set of concerns, instead of a fixed list of questions. With a focus, the observer has the flexibility to explore various angles that was not foreseeable.

The method offers the product designers an opportunity to be fully immersed in the user's work environment, and learn about users' interpretations to events and product functionalities, language, and the structure of their work activities. It provides product

designers an understanding of user work and usability; and further suggests generic principles of usability and work concepts that might become the initial frame work of new products (Wixon, Holtzblatt, Knox, 1990).

The strength of contextual inquiry lies in its ability to gather detailed description of user's work context, and user-centered descriptions of problems. The product designers might even have a chance to brainstorm possible solutions with the users during the observation process (Raven and Flanders, 1996). The data is not only able to point out the problem areas, but also help product designers understand why the problems were encountered and how to best resolve them. With the nature of the data gathered, contextual inquiry is best used in the early stages of development to help develop product design guidelines.

The disadvantage of contextual inquiry comes from the time it required of both the observer and the user. As mentioned before, , depending on the work that is being observed, the observation may take hours to months or even years to complete; it is a significant time investment to ask for.

OTHER FIELD RESEARCH METHODS

Besides contextual inquiry, there are a collection of different techniques and tools that study users, their tasks, and their work environment in the actual context of those environments.

Field observation (Hom, 2003) is a field research method that in which product develop team member visits the user at the user's work place, observe the user's work activities; collect artifacts or gather data about the physical traits that marks the work place by photographing, note taking, or sketches; and interview the user about their work. Although similar to contextual inquiry, when carefully planned, field observation can be done in only a few hours, and the user and the observer take on the traditional roles of interviewer and interviewee; they do not form any kind of partnerships.

Ethnographic interviewing (Katner, Sova, and Rosenbaum, 2003) is in essence an interview held at the user's work site. While contextual inquiry is focused on observing user's activities with inquiries to supplement the observations, ethnographic interviewing asks questions about use. It is an alternative to contextual inquiry when time constraint does not allow long durations of observations, or when observation is difficult to arrange (such as surgery). The reasoning behind this method is that being in the users' environment, in their daily setting and surrounded by familiar artifacts makes the discussion more concrete, and enables the interview to uncover deeper factors such as reminders, routines, work flows, and collaborations.

Field usability testing (Katner, Sova, and Rosenbaum, 2003) merges traditional laboratory usability testing method with the spirit of field research by conducting the usability testing sessions in the participants' environment using their own equipment. By doing so, the evaluators are able to see how the product fits into the user's environment, and how the surrounding affects product usage. The evaluators will be able to gather data on how users interact with the product, and also context-rich qualitative data about the target users.

SYSTEM INSPECTION METHODS

System inspection methods examine the product, instead of studying the user. The methods involve having usability experts or product designers inspect the product in an ad hoc manner, or following a guideline. Because the users are not involved and a lower number of evaluation sessions needed, the inspection techniques in general can be executed more quickly and less expensive than user studies (Mack and Nielsen, 1993). But this does not mean inspection methods would be replacing user study methods any time soon. Some usability experts have found that inspection methods find areas that require further testing, while user studies find areas for design changes (Savage, 1996). There are also chances that the problems identified by the inspectors are false positives that will not actually hinder product usability (Nielsen, 1992).

EXPERT REVIEW

Expert review is an informal method used by one or more expert usability professionals to evaluate a user interface. Molich and Jeffries (2003) commented on expert review as: “[t]he only thing you can say about it is that it doesn’t require users other than the reviewer(s).” The method relies on the insights experts are able to provide from their deeper knowledge in their respective fields.

Perkins, Belge, and Ehrlich (1997) asked a diverse group of experts to evaluate their system design by letting them use the system over an extended period of time, and discuss the system with them through email and face-to-face meetings. The experts include psychologists, artists, design practitioners, and researchers. They reported that in very little time and cost than a focus group, they were able to learn about the product in many different aspects.

As effective as expert review may be, its main focus is to evaluate if the product design may impede user’s performance of a task; it does not yield insights into user’s conceptual use model.

HEURISTIC EVALUATION

Heuristic evaluation is an informal system inspection method where a small group of evaluators are presented with an interface design and asked to judge whether each of its elements follows a set of established usability principles (Nielsen, 1992). The method is intended to be a “discount usability engineering” method (Nielsen, 1992) that provides a way to do a usability evaluation more quickly, and with less cost. Because of its “discount” nature, heuristic evaluation was found to be the most commonly used UEM in a survey to the practitioners (Rosenbaum, Rohn, and Humburg, 2000).

CONDUCTING A HEURISTIC EVALUATION

Heuristic evaluation can be performed by experts and non-experts. It is difficult to do a heuristic evaluation with a single evaluator; it is near impossible for one person to find all usability problems. Yet it has been shown that when there are multiple evaluators, each were able to find different usability problems, thus the effectiveness of the problem can be improved by having a group of evaluators. Usually, 4 or 5 evaluators are able to report near 70% of the usability problems; additional evaluators often are not able find much more additional problems (Nielsen and Molich, 1990; Nielsen, 1992; Nielsen, n.d.a).

The heuristics are not specific guidelines, but are more like general rules that describe common properties of good usability designs. There is no standard set of heuristics in which the system is evaluated with. Molich and Nielsen (Nielsen and Molich, 1990) developed a small list of heuristics with nine usability principles after years of experience in teaching and usability engineering consulting. Nielsen (n.d.b) refined the list based on a factor analysis of 249 usability problems and derived a set of 10 heuristics. Bastien and Scapin (1994) adapting heuristic evaluation with a set of ergonomic criteria. Many users also developed their own sets of heuristics (Nielsen, 1994). The evaluator is also allowed to consider any additional usability principles outside of the list that may be relevant to the specific design, or develop a domain specific heuristics for a certain kind of products (Nielsen, n.d.a).

The system in inspection does not need to be a finished product; it can be done with just a paper prototype (Nielsen, n.d.a), therefore allowing the evaluation to be performed in the early stage of development.

The evaluators are asked to inspect the system individually, in any way they like, and then compare and combine the problems they found to form a comprehensive list with reference to the usability principles the design violated. The identified problems can then be rated to allocate the resources needed to fix them, and to see if additional usability efforts are needed (Nielsen, n.d.c).

THE GOOD AND THE BAD

As mentioned in the previous paragraphs, the main advantage of heuristic evaluation is its ability to be done in a short period of time with limited resources. The method is also very flexible and does not require advanced planning; it could be carried out as soon as the group of evaluators is assembled and that there is a product or a prototype to evaluate. Heuristic evaluation has also proved to be highly effective in finding usability problems (Jeffries et al. 1991, Kantner and Rosenbaum, 1997). However, there are also several drawbacks.

The effectiveness depends largely on the evaluators' skill and experience. Though non-experts are able to perform the evaluation as well as experts, it is very likely that they would not be able to find as many usability problems as the experts. A "bad" evaluator is also more likely to miss the problems that a better evaluator did not pick up, thus lowering the aggregated count of problems found (Nielsen, 1992). The flexibility given to the evaluators, allowing them to inspect the system anyway they want also means a lack of support and structure to the inspection process (Law and Hvannberg, 2004). When the evaluators are not well informed about the product domain, the inspection may be not as effective.

It is important to have a group of evaluators do the inspection. Past studies have shown that even with expert evaluators, a single heuristic evaluation was consistently yielding the worst results among different UEMS, while the collective results of group heuristic evaluations out perform them (Jeffries and Desurvire, 1992).

The structure of the method limits the findings to the violation of heuristics, the result does not provide direct suggestion on how to improve the evaluated design, or lead to breakthroughs in the evaluated design (Nielsen and Molich, 1990). Neither does it

provide user data or add understandings to the context of a real user task (Simeral and Branaghan, 1997). It is also questionable whether the problems identified are in fact usability problems to real users. As Sears (1997) and Nielsen (1992) pointed out, the evaluators might identify every heuristic violation, such as minor, cosmetic issues, as a usability problem, although it does not hinder the usability performance. On the other hand, usability problems that are not violations to the heuristics may be missed.

It takes a larger group of evaluators to cover all of the usability problems. While 5 evaluators are able to identify the majority of usability problems, the margin of problems identified per additional evaluator diminishes to near none after that. It would require much more evaluators to find the rest of the usability problems. Law and Hvannberg (2004) found that they needed 16 evaluators to discover 75% of the usability problems in their study.

Expert opinions may not have as much impact to the designer team as seeing a user go through a usability test, therefore the influence to the final product design not as high (Kantnara and Rosenbaum, 1997).

ALTERNATIVE METHODS

With the promise of a cost effective and timely UEM, researchers have come up with alternative heuristic evaluation methods that seek to overcome its shortcomings.

Participatory Heuristic Evaluation (Muller et al. 1998) extends heuristic evaluation in an attempt to combine usability with the concepts of participatory design. The method adds several process-oriented principles that concerns the users' work processes to the originally product oriented heuristics. It also calls for users or user work-domain experts to be a part of the evaluator team, so that the users are represented in the structured discussions.

Metaphors of human thinking (Hornbaek and Frokjaer, 2004) is a method that focuses the inspection on users' mental activity by using a set of metaphors inspired by classical introspective psychology. The goal is to see if certain important aspects of human thinking were factored into the user interface design. The procedure is similar to heuristic evaluations in that it also asks the evaluator to inspect the system with a list of principles, in this case metaphors. The evaluator must first identify three tasks that users usually do. The evaluator then performs the tasks, identify the major problems encountered during the process, and use the metaphors to find usability problems. Next, the evaluator does the task again, this time, taking the perspective of each metaphors, one at a time, work through the tasks, and identify usability problems using the metaphors along the way. Hornbaek and Frokjaer (2004) found that using metaphors of human thinking, the evaluators were able to find more severe usability problems than when using heuristic evaluation. The results of the findings are also more consistent; evaluators agree on the problems found. The method requires less time to perform, but is more difficult to learn.

For heuristic walkthrough, Sears (1997) combines the benefits of heuristic evaluations, cognitive walkthroughs, and usability walkthroughs by asking the evaluators to explore a prioritized list of tasks; afterwards, they explore the system and look for usability

problems using a list of usability heuristics as guidance. Evaluators then compare notes and assign a single rating to the usability problems they found.

COGNITIVE WALKTHROUGHS

Cognitive walkthrough (Lewis et al., 1990; Wharton et al., 1992; Rieman, Franzke and Redmiles, 1995) is a theoretically structured usability evaluation process that focuses on a user's cognitive activities, especially while performing a task. It can be carried out by individuals or groups, software developers or usability specialists, and on finished products or paper prototypes.

Based on a theory of exploratory learning and corresponding interface design guidelines, cognitive walkthrough is a task-based methodology that centers an evaluator's attention on the user's goals and actions during a task, and on whether the system design supports or hinders the effective accomplishment of those goals. Moreover, it is a form-based evaluation methodology in which relies on a set of forms to guide the evaluation process.

The theory behind the method describes human-computer interaction in four steps: the user sets a goal to be accomplished with the system, the user searches the interface for action options, the user selects the action that seems to make progress towards the goal, and finally the user performs the action and evaluates the system feedback.

Built upon the theory, cognitive evaluation begins with the designer or designer team specifying a series of tasks that users are likely to use the evaluated system to accomplish (build a graph), the goals behind each task (communicate through visual presentation), and the underlying sub-goals (start a graphic program, input data, use the to build a pie chart). The sequences of user actions to complete the tasks are identified; there may be multiple sequences for one task. The design team then take on the perspective of a user, work through the tasks, and fill out the forms. The forms play a major part of the process; they guide the evaluators' actions and explorations with a series of questions regarding actions taken to complete a task. Each form asks the evaluators to identify the immediate goal of an action; inspect the atomic actions that precedes and follows, and if they were well supported; scrutinize user's cognitive process; evaluate user's action options; examine if the identified goal can be achieved and if there are other options; and finally, observe the appropriateness of system feedback. Since a form is required for each individual action, copies of it may be filled out dozens of times through a completed walkthrough.

THE GOOD AND THE BAD

Cognitive walkthrough has shown to be an effective UEM (Lewis et al., 1990). It also provided an option for evaluating a system in early development with relatively lower cost. But the details of the procedure created difficulties in its execution.

The walkthrough methodology presupposes knowledge of cognitive science terms, concepts, and skills from the evaluators (Wharton et al., 1992). A lack of familiarity with the terminologies in the form, such as the definitions of goal and action, could lead to misunderstandings and affect the outcome.

At least one evaluator needs to be familiar with the concepts of the cognitive walkthrough theory, and the cognitive science terminologies used during the process in order for the walkthrough to be effective. Lewis et al. (1990) conducted cognitive walkthrough with

four evaluators, three of which have deep understandings of the core principles of the theory. Throughout the walkthrough, there was a high level of agreement among the three evaluators, but less with the fourth. The fourth evaluator also found fewer errors than the other evaluators.

Wharton et al (1992) observed that the form filling procedure can be quite tedious, enough so that it discouraged some evaluators to use the method. They suggest rotating the recordkeeping duty in the team to alleviate the situation.

The process can be lengthy and time consuming not only because of its detail oriented nature, but because of the evaluators who are also the system designers may be sidetracked by the identified issues and attempt to find a solution for them during the evaluation (Wharton et al. 1992). Evaluators may also find problems that are not related to the current task but need noting. In these events, following a looser approach may help the evaluation to be more successful and more satisfying to the evaluators (Wharton et al. 1992).

The focus on the identified tasks and individual task actions may cause set effects, in which the evaluators are so focused on one set of problems they were unable to recognize issues outside of the task. In other words, the focus on tasks may cause the evaluators to miss the big picture. The technique can sometimes lead evaluators to suboptimal solutions on the task level instead of finding a solution with the perspective of the whole system design (Wharton et al., 1992). The narrow focus can even lead evaluators to propose erroneous solutions that might benefit an atomic action but hinder other tasks.

As with other task-based system inspection methods, the effectiveness of the technique relies heavily on the task selection, and the experience and skills of the evaluator. The evaluation may be biased by how the tasks were initially defined. With no real users involved, it is also questionable whether the action sets fully represent the activities a users may engage in (Wharton et al., 1992). Furthermore, the set actions eliminate the natural exploration process users experience when using a system (Sears, 1997). The ways users actually use an application can be really surprising to the evaluators and designers at times.

EVALUATOR EFFECT

Evaluator effect is a factor in UEMs that requires further investigation. It is defined by Hertzum and Jacobsen (2003) as the differences in evaluators' problem detection and severity ratings. In their study, they found that the average agreement between any two evaluators range from 5% to 65%, and none of the UEMs is consistently better than the others. They believe the cause of this effect is that usability evaluation is a cognitive process that requires the evaluators to interpret what they see and find. The phenomenon was seen in informal UEM such as heuristic evaluation, as well as the strict procedural cognitive walkthrough, and the formal usability testing. It casts questions to whether UEMs produce valid and reliable results, or if the consistency of an evaluator can be trusted.

Hertzum and Jacobsen suggested practitioners to reduce evaluator effect by: involving an extra evaluator, especially in critical evaluations, to cover as much ground as possible; be

explicit on the goal of the evaluation and task selection; and finally, reflect on the evaluation procedures and establish usability problem criteria.

CONCLUSION

This paper gave an overview of four user study methods (surveys, focus groups, usability testing, and contextual inquiry), three system inspection methods (expert review, heuristic evaluation, and cognitive walkthrough), and looked at the evaluator effect that is hared by the UEMs.

It is difficult to say which UEM is better or more effective than others. Studies have been done to compare the thoroughness (how many of the total usability problems were found), validity (would the problems identified really cause usability problems in real life usage), reliability (are the results consistent), and effectiveness of two or more UEMs (Jeffries et al., 1991; Jeffries and Desurvire, 1992; Karat, Campbell, and Fiegel, 1992; Kantner and Rosenbaum, 1997; Simeral and Branaghan, 1997). But the study by Hartson, Andre, and Williges (2001) points out that without a standardized measurement and a loose definition of the measurement criterions used in the studies, these comparisons should not be treated as final conclusions. And with the UEMs having far too many differences to be presented by independent variables, putting them side by side would be comparing apples to oranges. The only thing that the UEMs have in common is that they produce usability problem lists when applied to a target design. One would have to choose a UEM based on the conditions under which the method will be performed. And as Jeffries and Desurvire (1992) indicated, there are limitations to different UEM, especially the inspection methods; to fully evaluate product usability, one must have a more balanced repertoire of usability techniques.

REFERENCE

- Bastien, C. J.M and Scapin, D. L. (1995) Evaluation a user interface with ergonomic criteria. *International Journal of Human-Computer Interaction*, vol. 7, issue 2, 103-134.
- Beyer, H. R. and Holtzblatt, K. (May, 1995). Apprenticing with the customer. *Communications of the ACM*, vol. 38, issue 5, 45-52. Retrieved on 10/1/05, from ACM Portal.
- Boren, M. T. and Ramey, J. (2000). Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication*, vol. 43, no. 3, 261-278.
<http://www.idemployee.id.tue.nl/m.m.bekker/thinkaloud.pdf>. Last accessed, 11/11/05.
- Butler, K. A. (1996). Usability engineering turns 10. *interactions*, vol. 3, no. 1, 58-75. Retrieved on 9/13/05, from ACM Portal.
- Butler, M. B. 1996. Getting to know your users: usability roundtables at Lotus Development. *Interactions*, vol. 3, no. 1, 23-30. Retrieved on 11/11/05, from ACM Portal.
- Dicks, R. S. 2002. Mis-usability: on the uses and misuses of usability testing. In *Proceedings of the 20th Annual international Conference on Computer Documentation* (Toronto, Ontario, Canada, October 20 - 23, 2002). SIGDOC '02. ACM Press, New York, NY, 26-30. Retrieved on 9/30/05, from ACM Portal.
- Dumas, J. S. (2003). User-based evaluations. In *The Human-Computer Interaction Handbook*. J. K. Jacko, and A. Sears ed. Manwah, NJ: Lawrence Erlbaum Associates, Inc.
- Dumas, J. S., and J. C. Redish (1993). *A practical guide to usability testing*. Norwood, NJ: Ablex Publishing Corporation.
- Hom, J. (1998). Ethnographic study/Field observation. *The Usability Methods Toolbox*. Last accessed 11/12/05 at <http://jthom.best.vwh.net/usability/usable.htm>.
- Hornbaek, K., Frokjaer, E. (2004). Usability inspection by metaphors of human thinking compared to heuristic evaluation. *International Journal of Human-Computer Interaction*, vol. 17, no. 3, 357-374. Retrieved on 10/17/05, from Academic Premier.
- Jeffries, R. and Desurvire, H. (1992). Usability testing vs. heuristic evaluation: was there a contest? *SIGCHI Bulletin*, vol. 24, no. 4, 39-41. Retrieved on 9/30/05, from ACM Portal.
- Jeffries, R., Miller, J. R., Wharton, C., and Uyeda, K. 1991. User interface evaluation in the real world: a comparison of four techniques. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Reaching Through Technology* (New Orleans, Louisiana, United States, April 27 - May 02, 1991). S. P. Robertson, G. M. Olson, and J. S. Olson, Eds. CHI '91. ACM Press, New York, NY, 119-124. Retrieved on 9/30/05, from ACM Portal.
- Kantner, L., Sova, D. H., and Rosenbaum, S. (2003). Alternative methods for field usability research. In *Proceedings of the 21st Annual international Conference on Documentation* (San Francisco, CA, USA, October 12 - 15, 2003). SIGDOC '03. ACM Press, New York, NY, 68-72. Retrieved on 9/30/05, from ACM Portal.
- Kantner, L. and Rosenbaum, S. 1997. Usability studies of WWW sites: heuristic evaluation vs. laboratory testing. In *Proceedings of the 15th Annual international Conference on Computer Documentation* (Salt Lake City, Utah, United States, October 19 - 22, 1997). SIGDOC '97. ACM Press, New York, NY, 153-160. Retrieved on 9/30/05, from ACM Portal.
- Karat, C., Campbell, R., and Fiegel, T. 1992. Comparison of empirical testing and walkthrough methods in user interface evaluation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Monterey, California, United States, May 03 - 07, 1992). P.

- Bauersfeld, J. Bennett, and G. Lynch, Eds. CHI '92. ACM Press, New York, NY, 397-404. Retrieved on 9/30/05, from ACM Portal.
- Hartson, H. R., Andre, T. S., Williges, R. C. (2001). Criteria for evaluating usability evaluation methods. *International Journal of Human-Computer Interaction*, vol. 13, no. 4, 373-410. Retrieved on 10/17/05, from Academic Search Premier.
- Hertzum, M., Jacobsen, NE. (2003). The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction*, vol. 15, no. 1, 183-204. Retrieved on 10/15/05, from Academic Search Premier.
- Krug, S. (2000). *Don't Make Me Think! A Common Sense Approach to Web Usability*. Indianapolis, Indiana: New Riders Publishing.
- Kuniavsky, M. (2003). *Observing the User Experience: A Practitioner's Guide to User Research*. Burlington, MA: Morgan Kaufmann Publishing. Available on Books24x7.
- Law, E. L. and Hvannberg, E. T. 2004. Analysis of strategies for improving and estimating the effectiveness of heuristic evaluation. In *Proceedings of the Third Nordic Conference on Human-Computer interaction* (Tampere, Finland, October 23 - 27, 2004). NordiCHI '04, vol. 82. ACM Press, New York, NY, 241-250. Retrieved on 10/15/05, from ACM Portal.
- Lewis, C., Polson, P., Wharton, C., Rieman, J. (1990). Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Empowering People* (Seattle, Washington, United States, April 01 - 05, 1990). J. C. Chew and J. Whiteside, Eds. CHI '90. ACM Press, New York, NY, 235-242. Retrieved on 9/30/05, from ACM Portal.
- Lewis, J. R. (1994) Sample sizes for usability studies: additional considerations. *Human Factors*, vol. 36, no. 2, 368-378. Retrieved on 10/1/05, from Expanded Academic ASAP.
- Mack, R. and Nielsen, J. (1993). Usability inspection methods: report on a workshop held at CHI'92, Monterey, CA, May 3-4, 1992. *SIGCHI Bulletin*, vol. 25, no. 1, 28-33. Retrieved on 9/30/05, from ACM Portal.
- Melkus, L. A. (1985). The benefits of laboratory testing for usability. In *Proceedings of the Twenty-First Annual Conference on Computer Personnel Research* (Minneapolis, Minnesota, United States, May 02 - 03, 1985). J. C. Wetherbe, Ed. SIGCPR '85. ACM Press, New York, NY, 91-96. Retrieved on 10/1/05, from ACM Portal..
- Molich, R. and Jeffries, R. (2003). Comparative expert reviews. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA, April 05 - 10, 2003). CHI '03. ACM Press, New York, NY, 1060-1061. Retrieved on 10/1/05, from ACM Portal..
- Muller, M. J., Matheson, L., Page, C., and Gallup, R. (1998). Methods & tools: participatory heuristic evaluation. *Interactions*, vol. 5, no.5, 13-18. Retrieved on 10/23/05, from ACM Portal.
- Nielsen, J. (n.d.a). How to conduct a heuristic evaluation. *Useit.com*. Last accessed on 11/13/05, at http://www.useit.com/papers/heuristic/heuristic_evaluation.html.
- Nielsen, J. (n.d.b). Ten usability heuristics. *Useit.com*. Last accessed on 11/13/05, at http://www.useit.com/papers/heuristic/heuristic_list.html.
- Nielsen, J. (n.d.c). Severity ratings for usability problems. *Useit.com*. Last accessed on 11/13/05, at <http://www.useit.com/papers/heuristic/severityrating.html>.
- Nielsen, J. (1992). Finding usability problems through heuristic evaluation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Monterey, California, United

- States, May 03 - 07, 1992). P. Bauersfeld, J. Bennett, and G. Lynch, Eds. CHI '92. ACM Press, New York, NY, 373-380. Retrieved on 9/30/05, from ACM Portal.
- Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Celebrating interdependence* (Boston, Massachusetts, United States, April 24 - 28, 1994). B. Adelson, S. Dumais, and J. Olson, Eds. CHI '94. ACM Press, New York, NY, 152-258. Retrieved on 9/30/05, from ACM Portal.
- Nielsen, J. (1994). Guerrilla HCI: Using discount usability engineering to penetrate the intimidation barrier. Available at http://www.useit.com/papers/guerrilla_hci.html. Last accessed: 11/11/05.
- Nielsen, J. (1995). Usability inspection methods. In *Conference Companion on Human Factors in Computing Systems* (Denver, Colorado, United States, May 07 - 11, 1995). I. Katz, R. Mack, and L. Marks, Eds. CHI '95. ACM Press, New York, NY, 377-378. Retrieved on 9/30/05, from ACM Portal.
- Nielsen, J. (Dec. 1999). Voodoo usability. *Alertbox*. Last accessed on 11/11/05 at <http://www.useit.com/alertbox/991212.html>.
- Nielsen, J. (Aug. 2003). Usability 101: Introduction to usability. *Alertbox*. Last accessed on 10/24/05, at <http://www.useit.com/alertbox/20030825.html>.
- Nielsen, J. (2005). Ten usability heuristics. *Useit.com*. Last accessed on 11/14/05, at http://www.useit.com/papers/heuristic/heuristic_list.html.
- Nielsen J. and Landauer, T. K. (1993). A mathematical model of the finding of usability problems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Amsterdam, The Netherlands, April 24 - 29, 1993). CHI '93. ACM Press, New York, NY, 206-213. Retrieved on 10/13/05, from ACM Portal.
- Nielsen, J. and Molich, R (1990). Heuristic Evaluation of User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Empowering People* (Seattle, Washington, United States, April 01 - 05, 1990). J. C. Chew and J. Whiteside, Eds. CHI '90. ACM Press, New York, NY, 249-256. Retrieved on 10/10/05, from ACM Portal.
- Perkins, R., Belge M., Ehrlich, K. (1997). Expert reviews: Design for rapidly changing times. *interactions*, vol. 4, no. 3, 23-30. Retrieved on 10/17/05, from ACM Portal..
- Raven, M. E. and Flanders, A. (1996). Using contextual inquiry to learn about your audiences. *SIGDOC Asterisk Journal of Computer Documentation*, vol. 20, no. 1, 1-13. Retrieved on 10/1/05, from ACM Portal.
- Rieman, J., Franzke, M., and Redmiles, D. (1995). Usability evaluation with the cognitive walkthrough. In *Conference Companion on Human Factors in Computing Systems*. I. Katz, R. Mack, and L. Marks, Eds. CHI '95. ACM Press, New York, NY, 387-388. Retrieved on 9/30/05, from ACM Portal.
- Rosenbaum, S., Cockton, G., Coyne, K., Muller, M., and Rauch, T. (2002). Focus groups in HCI: wealth of information or waste of resources? In *CHI '02 Extended Abstracts on Human Factors in Computing Systems* (Minneapolis, Minnesota, USA, April 20 - 25, 2002). CHI '02. ACM Press, New York, NY, 702-703. Retrieved on 11/11/05, from ACM Portal.
- Rosenbaum, S., Rohn, J. A., and Humburg, J. (2000). A toolkit for strategic usability: results from workshops, panels, and surveys. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (The Hague, The Netherlands, April 01 - 06, 2000). CHI '00. ACM Press, New York, NY. Retrieved on 9/30/05, from ACM Portal.

- Rowley, D. E. (1994). Usability testing in the field: bringing the laboratory to the user. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Celebrating interdependence* (Boston, Massachusetts, United States, April 24 - 28, 1994). B. Adelson, S. Dumais, and J. Olson, Eds. CHI '94. ACM Press, New York, NY, 252-257. Retrieved on 9/30/05, from ACM Portal.
- Sato, S. and Salvador, T. (1999). Methods & tools: Playacting and focus troupes: theater techniques for creating quick, intense, immersive, and engaging focus group sessions. *interactions* vol. 6, no. 5, 35-41. Retrieved on 11/11/05, from ACM Portal.
- Savage, P. (1996). User interface evaluation in an iterative design process: a comparison of three techniques. In *Conference Companion on Human Factors in Computing Systems: Common Ground* (Vancouver, British Columbia, Canada, April 13 - 18, 1996). M. J. Tauber, Ed. CHI '96. ACM Press, New York, NY, 307-308. Retrieved on 10/1/05, from ACM Portal.
- Sears, A. (1997). Heuristic walkthroughs: Finding the problems without the noise. *International Journal of Human-Computer Interaction*, vol. 9, no. 3, 213-234. Retrieved on 10/10/05, from Academic Search Premier.
- Simeral, E. J. and Branaghan, R. J. (1997). A comparative analysis of heuristic and usability evaluation methods. In the proceedings of *Society of Technical Communication 44th Annual Conference*, 307-309. Last accessed on 11/11/05, at: <http://tc.eserver.org/20162.html>.
- Virzi, R. A. (1990). Streamlining the design process: Running fewer subjects. *Proceedings of the Human Factors Society, 34th Annual Meeting*, 291-294. Santa Monica, CA: Human Factors and Ergonomics Society. Retrieved on 10/1/05, from ACM Portal.
- Wilson, C. (1998). Pros and cons of co-participation in usability studies. *Usability Interface*, vol. 4, no. 4. Last accessed on 11/12/05, at <http://www.stcsig.org/usability/newsletter/9804-coparticipation.html>.
- Wixon, D., Holtzblatt, K., and Knox, S. (1990). Contextual design: an emergent view of system design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Empowering People* (Seattle, Washington, United States, April 01 - 05, 1990). J. C. Chew and J. Whiteside, Eds. CHI '90. ACM Press, New York, NY, 329-336. Retrieved on 10/1/05, from ACM Portal.
- Wixon, D. R., Ramey, J., Holtzblatt, K., Beyer, H., Hackos, J., Rosenbaum, S., Page, C., Laakso, S. A., and Laakso, K. (2002). Usability in practice: field methods evolution and revolution. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems* (Minneapolis, Minnesota, USA, April 20 - 25, 2002). CHI '02. ACM Press, New York, NY, 880-884. Retrieved on 10/1/05, from ACM Portal.
- Wharton, C., Bradford, J., Jeffries, R., Franzke, M. (1992). Applying cognitive walkthroughs to more complex user interfaces: experiences, issues, and recommendations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Monterey, California, United States, May 03 - 07, 1992). P. Bauersfeld, J. Bennett, and G. Lynch, Eds. CHI '92. ACM Press, New York, NY, 381-388. Retrieved on 9/30/05, from ACM Portal.