EVALUATING DISTRIBUTED USABILITY: THE ROLE OF USER INTERFACES IN AN ACTIVITY SYSTEM

Lejla Vrazalic
Information Systems
University of Wollongong
Wollongong NSW 2522
lejla@uow.edu.au

ABSTRACT

Traditional definitions of usability localise this fundamental Human Computer Interaction (HCI) concept in the user interface and reduce it to a variety of qualitative and quantitative attributes of the computer system. This simplistic view of usability has been used as the basis for developing design and evaluation methods in the discipline. This paper argues that, as a result, HCI methods are ineffective and suffer from various shortcomings. It is proposed that the notion of usability must be extended to include contextual factors, and viewed as being distributed across an activity system. Adopting this notion of distributed usability then requires a review of existing HCI methods. Usability testing, as a complete and self-contained HCI method, was chosen for this purpose, and the result, a distributed usability evaluation method (DUEM), is presented in this paper.

INTRODUCTION

In his article, *Trouble in Paradise: Problems Facing the Usability Community*, Rhodes (2000) outlines a gloomy future for usability by stating "usability as we know it is dying". He argues that usability is outdated, misunderstood and faces serious challenges in the face of emerging web technologies because new usability ideas, techniques and methods are not being developed. Rhodes (2000) arguments are compelling, but his observations are not new. The Human Computer Interaction (HCI) and usability communities are being faced with mounting and pressing concerns for which an instantaneous remedy is not readily available. It is the premise of this paper that in order to begin resolving these concerns, it is necessary to reflect on the very fundamental concept that the discipline is based on – the concept of usability, and then examine the implications of this on HCI methods, techniques and tools.

This paper argues that there are intrinsic problems with our current definition of usability and that it is necessary to redefine usability to encompass more than just the computer system in isolation. Spinuzzi's (1999) notion of distributed usability and Nardi and O'Day's (1999) information ecology are proposed as a means of doing this. The implications of extending the definition of usability are significant to the way in which existing HCI methods are viewed and used. Usability testing, as a complete and self-contained HCI method, has been chosen to demonstrate these implications. A case will be made that the problems associated with current usability testing are symptomatic of the traditional view of usability and can be overcome by developing a testing method based on distributed usability. Activity Theory, and its associated principles, will be proposed as the underlying framework for developing this method. A brief description of the method will be provided with an example of its application to demonstrate its benefits. An operational model of the method is currently being finalised. For the purposes of this discussion, the terms "computer system", "system" and "interface" will be used interchangeably.

TRADITIONAL USABILITY

> "*Many developers dream of an algorithm giving an exact measure of the usability of a product. Ideally, one could take the source code of a program, run it through an analysis program giving a single number in return: Usability = 4.67*" (SINTEF Group, 2002)

HCI as a discipline bases itself on the need to design, evaluate and implement computer systems for human use (ACM SIGCHI, 1992). The notion of usability is therefore fundamental to HCI. Usability is an abstract concept intended to encase and encompass both the design and evaluation of computer systems. It is the glue that binds the entire systems design and development process together. To some, usability is a science. To others, it is an art form. As such, usability is a concept that does not lend itself to a precise and clear-cut definition. Generally, usability refers to the ease of operating a system interface. The simplicity of this statement may appear to be misleading due to the plethora of other, seemingly more comprehensive, definitions of usability that exist. However, most of these definitions are based on this central notion of "ease of use" (Miller, 1971 cited in Shackel, 1986). Since the notion of ease of use in itself is somewhat vague, some authors and researchers have defined usability in terms of multiple high-level criteria or attributes.

Shackel (1986) proposes that usability can be specified and measured numerically in terms of four operational criteria: effectiveness (a required level of performance by a percentage of specific users within a range of usage environments), learnability (a pre-defined time period from the start of user training and based on a specified amount of training), flexibility (the levels of adaptation and variability in possible tasks) and attitude (user satisfaction levels after continued use). Nielsen (1993) views usability as a narrow concern when compared to the issue of system acceptability, and models usability as an attribute of system acceptability. However, similarly to Shackel (1986), Nielsen argues, usability itself can be further broken down into five attributes: learnability (the system should be easy to learn), efficiency (the system should be efficient so that high levels of productivity are possible), memorability (the system should be easy to remember and not require re-learning), errors (the system should have low error rate and enable quick recovery after errors) and satisfaction (the system should be pleasant to use).

In contrast to these high-level criteria and attributes, Norman (1988) conceptualises usability in terms of design principles based on a combination of psychological theory and everyday user experiences. The aim of these principles is to help designers make improvements to the system and explain different aspects of their design to the various stakeholders (Thimbleby, 1990). While numerous principles have been operationalised in HCI, the most well-known have been proposed by Norman (1988) and include: visibility (making the system functions visible so that each function corresponds with a control), mapping (a direct, natural relationship between the controls and their functions), affordance (perceived and actual properties of an object that determine how it can be used), constraints (limiting the behaviours and possible operations on an object) and feedback (sending back information to the user about what action has been done).

What is clear from the above is that a single universal definition for usability does not exist, which makes it a confusing concept to explain and, more importantly, justify to the business community (Rhodes, 2000). However, the general implication in most definitions of usability is that usability is located within the system itself. It can be thought of purely as an 'attribute of the entire package that makes up a system' (Dumas and Redish, 1993) and reduced to factors such as ease of use (*of the interface*), learnability (*of the interface*), memorability (*of the interface*), visibility (*of the interface*), mapping (*of the interface*) or any of the other attributes and principles discussed above. Shackel (1986) was one of the first researchers to alert the HCI community to need for an extended view of usability, embracing four principal components (the user, the task, the system and the environment). Usability, Shackel argued, was about achieving harmony in the interaction between these four components. Since then, a number of other researchers have increasingly become critical of the "traditional" localised view of usability (Thomas and Macredie, 2002; Spinuzzi, 1999; Nardi and O'Day, 1999; Beyer & Holtzblatt, 1998; Nardi, 1996; Engeström and Middleton, 1996; Hutchins, 1995; Kling & Iacono, 1989) because it does not take into consideration those attributes which extend beyond the computer system, including the social context, the work practices and the historical development of the activities that the computer system supports. Thomas and Macredie (2002) went so far in their criticism of traditional usability as to argue that the current conceptions of usability are ill-suited, unwieldy, meaningless and unable to handle the "digital consumer" (p 70). As a result, an alternative notion of distributed usability has emerged.

DISTRIBUTED USABILITY

Spinuzzi (1999) argues that the traditional view of usability is inadequate because it disregards the influences and consequences of contextual factors such as the interaction between humans, the use of artifacts other than the system and the actual work practices of users. Karat (1997) supports this view by describing usability as an attribute of the interaction with a system in a context of use. In Spinuzzi's (1999) opinion, usability is distributed across an activity network which is comprised of assorted genres, practices, uses and goals of a given activity. An activity network represents a unit of analysis that takes into account individual users working with others as part of a larger activity. Thinking about usability in this way provides us with a more encompassing view of the system (and its interface), the users, as well as the users' goals, and, Spinuzzi argues, leads us to consider solutions that we may not have if we had studied individuals alone. For example, in his study, Spinuzzi found that breakdowns in the users interaction with the system (as defined by Bødker, 1991) were not caused simply by the size of the mouse pointer or even the levels of user training, but could be attributed to "deeper discoordinations" between the interface and other genres relating to the context of use. This interpretation of usability would not have been possible if only localized attributes of the interface, such as ease of use, learnability and memorability, were examined.

Spinuzzi advocates the study of on-screen and off-screen genres, or typified forms, and their mediatory relationships in the context of an ecology of interrelated tools and activities. Nardi and O'Day (1999) also view this arrangement of tools, which jointly mediate activities, as belonging to an information ecology. They define an ecology as a "system of people, practices, values, and technologies in a particular local environment" (p.49) which focuses on human activities served by technology, rather than technology itself. Indeed, where traditional views of usability relate only to a system consisting of hardware and its associated software, distributed usability proposes to extend the system to include an ecological context consisting of people, their work practices and activities (supported by the hardware and software system as well as other *tools*) and the social context of these activities. This idea echoes Shackel's (1986) four principal components of usability (the user, the task, the system and the environment), however Spinuzzi (1999) and Nardi and O'Day (1999) view these components as being interrelated. The components cannot be fully understood individually because together they make up an extended system that is more than just a sum of its parts.

The distribution of usability across this extended system or ecology has significant implications for both the design and evaluation of systems. Clearly, the design process could not proceed without taking into account and devoting substantial resources to understanding the larger context of user activities, and incorporating this context into the design process. The implications for system evaluation, however, are even more considerable in light of existing evaluation methods, which rely on the traditional view of usability and therefore primarily focus on the system as an artifact in isolation and on assessing the localised attributes of that isolated system. If the HCI community is to adopt a distributed view of usability, it will be necessary to revisit our existing evaluation methods to determine their usefulness, reliability and validity, and develop new methods that are based on a fundamentally different notion of usability. To this end, a distributed usability evaluation method has been developed.

USABILITY TESTING

Usability testing is "the gathering of information about the use of prototypes of software products from users who are not involved in the design of the products themselves" (Holleran, 1991). In the 1980s, laboratory based usability testing, as described by Rubin (1994) and Dumas and Redish (1993) emerged as the golden standard for usability evaluation (Lewis, 2001). Since then, other usability evaluation methods have been developed, including heuristic evaluation (Nielsen & Molich, 1990) and cognitive walkthroughs (Polson et al, 1992). Unlike, usability testing, which involves observing users directly interact with a system, these methods are predictive and conducted by teams of experts. Bailey (1993) and Tullis (1993) have questioned experts' ability to predict usability problems following empirical studies, which indicated that experts experienced problems when trying to predict human performance. The results of these studies should come as no surprise to anyone who has been involved in usability testing with users. Gould and Lewis (1985) laid down the basic principles of user-centred design as being early focus on users and tasks, empirical measurement and iterative design. Usability testing is fundamental to the achievement of all three principles because it involves users, it enables empirical measurement to be undertaken, and it supports the notion of iterative design following testing of early system prototypes. Due to this central role of usability testing in user-centred design approaches, it was decided that usability testing, rather than predictive evaluation methods, would form the basis for developing a method to demonstrate the use of distributed usability. However, usability testing in its current form is plagued by practical and methodological problems. These will be discussed next.

Problems Associated with Usability Testing

Despite its status as the "golden standard" in usability evaluation, a number of authors have identified pitfalls associated with the usability testing process. Wixon and Wilson (1997), for example, have drawn attention to the problems with setting the usability testing goals during the initial stages of usability testing. They claim that usability testing goals may be too ambitious or too many in number. This in turn creates the need for more elaborate and lengthy test procedures, as well as generating more data to analyse. This is highly undesirable, considering that human factors experts spend on average 33.2 hours per user when conducting usability testing (Jeffries et al, 1991). Increasing the complexity of the testing process also increases the risk of failure (Wixon & Wilson, 1997). Conversely, arguments have been put forward that usability testing cannot be used to evaluate every aspect of the system (Wilson & Wixon, 1997), making the process deficient in collecting data about the whole system. Therefore, there appears to be a trade-off between the complexity of the test objectives and the need to test the system as a whole.

Holleran (1991), on the other hand, identified a number of methodological pitfalls in the usability testing process associated with sampling. He bases his argument that generalisability confidence increases with the number of participants who are representative of the target system users. While the former issue has been addressed by the work of Nielsen and Landauer (1993) and Virzi (1992), Holleran's (1991) concern lies in the degree to which the participants are representative of typical users. This issue is problematic for two reasons: participants may only be as representative as the evaluators' ability to understand and categorise the target user population (Rubin, 1994) and may be difficult to identify, access and recruit representative users, particularly where online systems are being tested.

Generating the scenarios to be used in the usability tests is another area of contention owing to the complexity of the task. Scenarios are intended to reflect the activities that typical users perform and their associated goals. However, in practice, most scenarios simply reflect functions and aspects of the system that are representative of its capabilities. The implication is that, rather than evaluating whether the system supports the users' tasks, usability testing is concerned with the evaluation of specific functions of the system that have been implemented, regardless of the usefulness or relevance of these functions. Generally, scenarios for usability testing are generated by the developers (Hartson et al, 2001) and/or evaluators which introduces an inherent bias into the evaluation. Jacobsen et al (1998) and Hertzum and Jacobsen (2001) have termed this bias the evaluator effect. The presence of the evaluator effect means that different evaluators who are testing the same system will detect substantially different sets of usability problems because the use of their judgement is required. Hertzum and Jacobsen (2001) claim that the evaluator effect "persists across differences in the system domain, system complexity, prototype fidelity, evaluator experience, problem severity, and with respect to detection of usability problems as well as assessments of problem severity" (p 439). Their findings show that on average the agreement between any two evaluators can range from 5% to 65% which is an astonishing discrepancy. The CUE studies by Molich et al (1998; 1999) have found even more divergent levels of agreement.

The involvement of evaluators and/or developers in the task scenario generation process also often results in the wrong terminology being used to describe users' tasks. It is critically important that scenarios are written in the language of the users because the way in which the users interpret the scenarios will ultimately have an effect on the results of the usability testing. If, for example, a user is unable to understand the requirements of the task, this indicates that the task is not representative and that the user may not be able to complete the tasks because of interpretation rather than problems with the system.

Evolving from formal experiments, usability testing has traditionally been situated in a laboratory setting. This controlled environment and the use of testing equipment such as recorders and cameras is designed to enable evaluators to evaluate a system as objectively as possible and generate quantitative data. This artificially created space, however, carries with it a series of negative connotations, with users known as *subjects*, *controlled* by the evaluators measuring specific *variables*. Despite the seeming objectivity enabled by the testing environment, there are also several subjective factors involved at crucial stages of the testing process, including during the setting of usability goals, the generation of task scenarios and the interpretation and analysis of results, both carried out by expert evaluators. While the quantitative performance data collected during a usability test may be useful in general terms, it may not be a reflection of actual performance because contextual influences that are inherent to real user activities have not been factored into the data. Finally, the use of descriptive and inferential statistics does not necessarily provide insights into whether a system actually works. Measures of statistical significance used in analysing quantitative data is simply a measure of probability that the results did not occur due to chance (Rubin, 1994). This type of 'micro-level' analysis does not actually prove that the system is usable, or more importantly, useful.

A major shortcoming of a laboratory based usability test is the unnaturalness and artificiality of the environment (Hartson et al, 2001; Wilson & Wixon, 1997; Rubin, 1994). Testing carried out in a laboratory is radically different to the natural, everyday practices that humans engage in through interaction with other humans, systems and tools. This reduces the ecological validity of the evaluation process. According to Thomas and Kellog (1989), ecological validity refers to how close a testing situation is to the real world. They identify four ecological gaps in laboratory based usability testing. The mismatch between the users' real context and a test context is referred to as a "work-context gap". The work-context gap does not only include differences in the physical context, but the job context and social and cultural contexts of user activities. Rubin (1994) also points out that any form of usability testing (in a laboratory or in the field) depicts only the situation of usage and not the situation itself. Naturally, this will have an effect on the test findings because the situation in this instance represents the context of the usage and the two are inextricably linked. Rubin (1994) has suggested creating laboratory spaces which resemble the users' real context, however, even with these improvements, usability testing is not a perfect indicator of field performance (Nielsen & Phillips, 1993). One of the reasons for this lies in the participants' motives. Holleran (1991) argues that participants in a usability test will persevere in doing tasks which they are unlikely to do in their own context out of willingness to comply with the evaluators' requirements. Thomas and Kellog (1989) have called this the "user gap". It is also more widely known as the Hawthorne effect and amplified by the presence and use of video cameras in a usability lab to record testing sessions. The artificiality of the laboratory is also conducive to the use of brief and clearly defined task scenarios which are usually completed within a specific time period. This is in stark contrast to the ill-defined and ongoing activities that users actually engage in. This mismatch is known as the "task gap". Finally, the "artifact gap" refers to the differences between short-term system usage during a test and long-term usage in the real world. Some usability problems may only emerge after prolonged system usage which is clearly not possible in the context of a laboratory.

Holleran (1991) also questions the validity and reliability of the data collected during usability testing. Validity refers to whether the evaluators are actually measuring what they intend to measure. Considering the problems associated with deriving representative task scenarios, the validity of usability testing results remains a controversial issue. Furthermore, it is not always possible to collect quantitative data or data for which suitable statistical measurements are possible (Holleran, 1991). Reliability is the extent to which the data produced in one usability test will match the data produced in another if the testing is replicated under the same conditions. The CUE studies (Molich, 1998; 1999) have shown the reliability of usability tests to be quite low. The end result of a usability test is a series of, presumably, usability problems that need to be fixed. Just as there is no clear definition of usability, the characterization of a usability problem remains elusive and can perhaps be seen as one of the "holy grails" of HCI. The term usability problem is commonly used to refer to any difficulties or trouble a user may have while using the system, or any faults in the system which cause a break down in the interaction. However, there is no explicit criteria defining when such a difficulty or fault constitutes a usability problem (Hertzum & Jacobsen, 2001) and so any problem reported is deemed to be a usability problem. Different evaluation methods identify different usability problems and no one single method can be relied on to uncover every usability problem with a system. In fact, it is quite possible for the same method to produce different outcomes as Molich et al. (1998, 1999) demonstrated in the CUE studies.

Usability testing is also plagued by logistical problems such as participants not showing up (Wilson & Wixon, 1997), scheduling convenient times for the testing, and problems with obtaining and maintaining specialised equipment. As such, usability testing is the most expensive and time consuming UEM (Hartson et al, 2001; Jeffries et al, 1991). It also requires a working prototype and specialised expertise to conduct (Jeffries et al, 1991) because, even though users take part in the process their role is restricted. Usability testing is strongly controlled by the evaluators at every stage (Mayhew, 1999) and driven by the system (Sweeney et al, 1993). The users are reduced to being passive participants who are controlled, observed, recorded and surveyed in order to collect performance data. The evaluators decide what, how, when and where to evaluate and the users have no input into the design of the evaluation or the interpretation of the results.

Finally, one of the most critical disadvantages of usability testing is that, like most other UEMs, it has no theoretical basis. Loosely based on the formal experiment, the method emerged from practice and has since been widely adopted and applied. Different approaches to usability testing are used and then compared to find out which approach works better and why, without any theoretical framework to allow this type of analysis. Holleran (1991) refers to this as "dustbin empiricism".

The problems with usability testing identified above are symptomatic of the narrow, traditional view of usability as a localised attribute or quality of a single system. Sweeney et al (1993) support this view by arguing that usability testing in a laboratory tends to be driven by the system. The use of a laboratory for the testing process enables evaluators to focus only on the system and eliminate any external or contextual variables that would interfere with the assessment of the system performance, thus making it easier to isolate and identify problems. The system and its functions drive the selection of users for the test sample because representative users are usually chosen *after* the functions of the system are known. The tasks developed for the usability test are also based on the system functions, rather than user needs. Instead of assessing whether the system does what the users want it to do, usability testing usually tests how well the system does what it can do. Finally, the confusion surrounding usability is reflected in the difficulty of defining a usability problem.

By adopting the notion of distributed usability as the starting point for the development of a usability testing method, it may be possible to eliminate the problems discussed above. However, a notion in itself is insufficient for this purpose. As mentioned above, it is necessary to support the notion with a theoretical framework on which to base the distributed usability testing method. Cultural Historical Activity Theory (CHAT) offers such a framework.

CULTURAL HISTORICAL ACTIVITY THEORY

Cultural Historical Activity Theory, or simply Activity Theory (AT) as it is widely known, provides a broad conceptual framework that can be applied to the human-computer interface in such a way as to empower the computer user with the necessary tools to work though the interface in order to achieve desired outcomes. Historically, AT draws on the Vygotskian (1978) theory of tool mediation or the mediation of human activities by the use of tools. This approach deviates from the cognitive approach in that the computer system is seen as distinctly different in both character and composition to its human user. From an AT perspective, people are embedded in a socio-cultural context and their behaviour cannot be understood independently of it. Furthermore they are not just surrounded by the context but actively interact with it and change it. Humans are continually changing activities and creating new tools. This complex interaction of individuals with their surroundings has been called an activity and is regarded theoretically as the fundamental unit of analysis, a system that has structure, its own internal transitions and transformations, its own development (Leont'ev, 1981).

AT is becoming more widely known by HCI researchers in the West (Kuttii, 1996; Engeström, 1995; Kaptelinin, 1994; Bødker, 1991; Nardi, 1996) since it was introduced in Russia in the eighties and early nineties. It's most current and widely-adopted form is Engeström's (1987) systemic model shown in Figure 1. In this model, the subject refers to the individual or group engaged in the activity, while the object refers to that (either 'raw material' or 'problem space') at which the activity is aimed. The object defines the activity and is transformed into an outcome using physical and symbolic mediating tools. The community consists of individuals who share the same object of the activity, while the division of labour refers to both the horizontal division of tasks between the members of this community and to the vertical division of power and status. The rules refer to the explicit and implicit regulations, norms and conventions constraining the interactions within the activity system (Engeström, 1987).

Kuutti (1996) describes the key principles of AT as follows:

Activity as the basic unit of analysis

Instead of analysing only human actions, AT proposes that a minimal meaningful context for these actions should be included in the analysis and this unit comprising actions in a context is an activity.

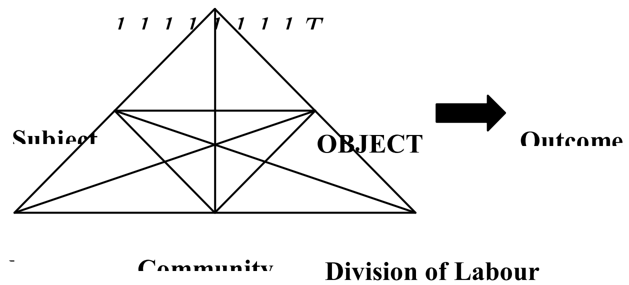

Figure 1. The Human Activity System (Engeström, 1987)

History and development

Activities are in a constant state of evolution and therefore, it is necessary to historically analyse an activity in order to gain an understanding of the current situation. A hierarchical system of contradictions is inherent to activity systems and it is these contradictions that cause an activity to develop over time. Engeström (1987) states that contradictions emerge as a result of conflicts within and between activity systems. He categorizes four levels of contradictions:

- Primary contradictions *within* the elements of the central activity, usually between the value and exchange value of an element;
- Secondary contradictions arise *between* the elements of the central activity;
- Tertiary contradictions take place between the object of the activity and the object of a more culturally advanced activity, and
- Quaternary contradictions occur between the activity and its 'neighbouring' activities, such as the tool-producing activity, subject-producing activity, etc.

Artifacts and mediation
Activities are mediated by artifacts and artifacts themselves are created during the development of an activity. This dual relationship further implies the developmental nature of activities.

Structure of an activity
An activity is directed towards an object and the object is what distinguishes one activity from another. The transformation of the object into the outcome motivates the existence of the activity. Furthermore, the object and motive could undergo changes during the development of an activity.

Levels of an activity
An activity, which is driven by motives, is realised through conscious actions which are directed towards specific goals. Those actions, in turn, are implemented through operations dependent on the available conditions. The relationship between the elements of this hierarchy, depicted in Figure 2, is dynamic. For example, when working with a computer for the first time, using a mouse is a conscious action requiring the deliberate attention of the user. Through practice, this action will collapse to the level of operations where it becomes habitual and subconscious. However, if the conditions change such that the mouse stops working, the user will be forced to focus his/her actions towards the mouse once again, returning the operation to the level of a conscious action. This type of interruption in the internal structure of an activity is termed a breakdown (Bødker, 1991). Breakdowns occur for various reasons, but they are evident in system use when the system itself becomes the object of the user's actions, rather than having a mediating role.

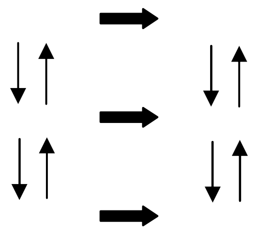

Figure 2. Structure of an Activity (Leont'ev, 1981)

The notion of distributed usability implies the distribution of usability across an activity network comprised of assorted genres, practices, uses and goals of a given activity. Activity Theory provides a unifying framework for these elements because the given activity is the basic unit of analysis consisting of the users (the subjects), their task (the object), the system and other objects utilised to complete the task (the mediating tools) and the social context (the community, the division of labour and the rules). Furthermore, the structure of an activity is concerned with the users' motives, their goals and the conditions in which these are made possible. A distributed usability testing method based on Activity Theory has been developed. The following section will briefly outline the operationalisation of this method.

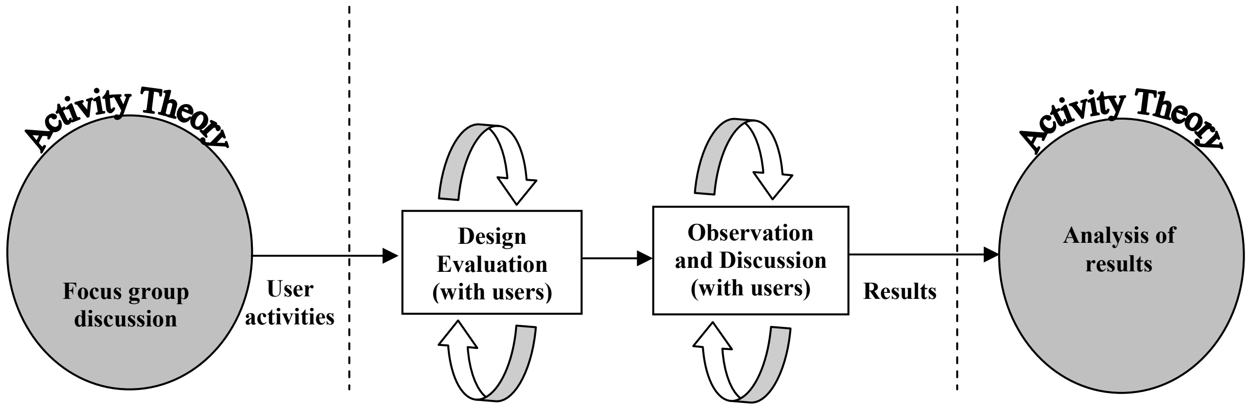

Figure 3. A model of the Distributed Usability Evaluation Method (DUEM)

DISTRIBUTED USABILITY EVALUATION METHOD

The Distributed Usability Evaluation Method (DUEM) places the users at the centre of the evaluation by involving them in the design of the evaluation process, and reduces the computer system to a support role as one of the many mediating tools in user activities. The focus, instead, is on identifying problems across the entire activity network. Specific usability attributes, such as ease of use, learnability and memorability are not examined or measured. The main concern is to identify usability problems which are caused by breakdowns in the interaction and contradictions in the activity network. A working model diagram of the method is shown in Figure 3.

Phase I: Understanding Users' Activities

The initial phase involves gaining an in-depth understanding of real user activities by observing and interviewing users who actually use the system being tested in their everyday activities. However, the system itself is not the central focus of this phase. It is only important to the extent to which it supports the users' activities. Where appropriate, field interviews and observations can be carried out in order to understand users' needs, desires and their approach to the work they do (Beyer and Holtzblatt, 1999). The interviews can be carried out on a one-to-one basis or in focus groups involving teams that carry out the same activity. This provides a forum for discussing and observing the social interactions between users, and for developing an understanding of the social context by gathering information, stories and anecdotes. Due to the problematic nature of gathering this type of ad hoc information, the AT principles described previously can be used to make sense of the information gathered and also provide evaluators with a common vocabulary (Nardi, 1996) as AT terminology is a close reflection of users' activities and, as such, easily understood by users. Questions based on AT such as the following can be used as a guide to understand an activity:

- "Who are the users?" – to collect information about the users' background in relation to three things: the general activities they perform which are supported by the system being evaluated; the system itself; and the use of the system specifically to carry out the activities. This provides evaluators with an insight into the users' knowledge of the domain (activity), the system (mediating tool) and the system use (actions).
- "What is the object of the users' activity?" – to appreciate the essence of what the users are engaged in. The object is the "ultimate goal" of the user. An understanding of the object allows evaluators to situate the system and its role in the activity. It draws the attention away from the system itself, and positions it in relation to what the user is trying to achieve.
- "What are the tools that the user makes use of?" – to determine the complete set of tools that a user makes use of in carrying out the activity. The system will constitute only one of

many tools, both material and psychological in nature. Since the use of the tools mediates the users' activities and different actions can involve the use of a different tools, the evaluators will gain an insight into the alternative ways in which the activity can be undertaken. If the users' indicate a preference for a specific set of actions over others to carry out an activity, this may have implications for the interface.

- "Who does the user interact with in carrying out the activity?" – to understand the users' community, the rules of this community and the division of labour within the community. This represents the users' social context and allows evaluators to identify a full set of system stakeholders.
- "How was the activity carried our prior to the system being evaluated?" – to define the historical development of the activity and any remnants of the previous state of the activity which may have an impact on its current form.

The information collected from the interviews or focus groups provides an integrated, holistic view of the main activity and other intersecting activities and a description of the various mediating tools used in performing the activity, as well as an explanation of how they are used. The key objective of this phase is to explore the users' work practice (Borgholm & Madsen, 1999) and gain an understanding of real user activities. It is important to allow the evaluators to immerse themselves in the users' practice and, by applying AT principles, gain a shared understanding and interpretation of what transpires during a typical activity which is supported by the system being evaluated. Once a common interpretation has been developed, the evaluators can proceed with phase two, which involves planning the evaluation process with the users.

Phase II: Evaluating with Users

Unlike with traditional usability testing, the involvement of users does not end with Phase I. Instead, users are directly involved in planning the evaluation process. During Phase II, the evaluators and users work jointly and iteratively to design and develop a set of evaluation goals based on the user activities identified during Phase I. The evaluation goals are not system specific. They are defined in relation to user activities. This is in contrast to traditional usability testing which uses the system as the starting point for developing evaluation goals. Based on these goals, a means of conducting the evaluation is collaboratively negotiated between the users and the evaluators. This can involve the use of activity scenarios or simply a free-form evaluation whereby users and evaluators use and discuss the interface in the users' natural environment. The latter form is preferable because it enables evaluators to study the impact of the social context. Since the DUEM does not focus on attributes of the system and micro-level analysis (such as the number of clicks, or time taken to complete a task) per se, the use of cameras and recording is not required.

It is important to note that during Phase I of the evaluation process, the primary focus was on user activities, and not the system. The system is discussed only to the extent of the role it plays in the users' activities in conjunction with other tools. In Phase II, this role is actually evaluated from the point of view of the activities. The use of prescriptive scenarios to test the functions of the system is not advocated. Instead, descriptive scenarios may be drawn up by users and evaluators based on the user activities as a guide to direct the evaluation process. The outcome of Phase II is qualitative data about the users' interaction with the system in the form of written notes based on observations and discussions which is then analysed using Activity Theory.

Phase III: Analysis with Users

During this phase, the rich data collected in Phase II is analysed in terms of the breakdowns and contradictions. The notes are used to identify breakdowns (as defined by Bødker (1991)) in the activity system and co-relate these breakdowns to the contradictions (as defined by Engeström (1987)). Breakdowns do not necessarily occur only in the actual use of the system. They can result from a change of focus in any element of the activity system but may have in impact on the system use. These breakdowns can be mapped to contradictions that are identified within and between activities. The mapping of breakdowns and contradictions defines the usability problems in the activity network. This process will be demonstrated with an example based on a system described by Nielsen (1990). The LYRE system, a French hypertext system used for teaching poetry, allows students to analyse poetry by adding new annotations to poems using hypertext anchors. The system was based on the French tradition of students working within a framework set up by the teacher, so it did not allow students to add new viewpoints, a facility reserved only for the teachers. In Scandinavian countries, the focus of teaching is on increasing students' potential to explore and learn independently. Had the LYRE system been implemented in Scandinavia, its use would have resulted in a series of breakdowns caused by primary contradictions within the tool (LYRE), secondary contradictions between the tool (LYRE), the subjects (the students and their expectations), the object (to explore and learn independently) and the division of labour (between the teacher and the students). Tertiary contradictions would also have occurred between the object of the French teaching activity (to analyse poetry) and the object of the Scandinavian activity (independent student exploration and learning). By identifying usability problems with the system as breakdowns caused by these contradictions, it would have been possible to determine whether the system supports the students' activities, instead of focusing on the ease of use, learnability or memorability of the system. It would have been possible to establish that the system is intrinsically flawed.

The DUEM proposes that even the analysis of the results be carried out collaboratively with the users. Their presence ensures that breakdowns in the activities can be clarified and considered without the need for re-interpretation by the evaluators. It also enables evaluators to assess which breakdowns are more serious in nature and discuss with users how these can be resolved. For example, during the evaluation of a web-based system of a large government organisation, a serious breakdown occurred which was mapped to a number of contradictions in the activity system. The system was developed to support investigative activities. In order to carry out those activities, users had to submit requests to supervisors using the system to access personal information about clients. Users were required to complete an online form which was then submitted to a supervisor for approval in the form of an e-mail message. However, this was observed to be ineffective in those cases where urgent approval was required. If, for example, a supervisor was away, the approval would be delayed (with serious consequences). To overcome this problem, users would first contact each supervisor by phone to determine who was available to approve a request immediately. They would then submit the request to that supervisor using the system. The breakdown in this activity occurred because the users had to shift their focus from the investigative activity to finding an available supervisor by phone. Although the users' direct interaction with the system did not cause this breakdown, a contradiction between the system (the tool), the object of the activity and the roles of stakeholders involved did, because the system was not designed to support the activity's division of labour effectively. When the analysis was carried out with the users, they provided two suggestions for fixing the problem: designing the system so that it alerts the users which supervisor is available (i.e. logged on to their e-mail account) and flagging the e-mail message so that supervisors were able to detect these types of requests easily and distinguish them from other e-mail messages. Had the above system been evaluated using the traditional usability testing method which focuses primarily on the interface, the usability problem would have remained undetected because there is no provision for evaluating the usability across the entire activity system. Since the problem was not directly related to the interface and a usability laboratory would not have afforded observing and understanding the way in which the actual activity was carried out, a serious flaw in the system may have been omitted.

DUEM offers a number of advantages to both researchers and practitioners. Primarily, the method overcomes most of the problems associated with traditional usability testing described previously, by involving users in the design of the evaluation process, and focusing user activities instead of system functions. DUEM does not evaluate the system in isolation from other elements in the activity network. This situation is only made possible by starting off with a view to distributed usability in the first place.

CONCLUSION

This paper has argued that fundamental problems with our current understanding of usability have emerged and that it is necessary to redefine usability to encompass more than just the computer system in isolation. The notion of distributed usability has been proposed as a means of achieving this. However, the implications of redefining this basic concept are significant to our existing usability evaluation methods, and usability testing in particular owing to its role in user-centred design. It has been argued that the problems associated with the current usability testing method are symptomatic of the traditional view of usability and can be overcome by developing an evaluation method based on distributed usability. Activity Theory, and its associated principles, has been proposed as the underlying framework for the Distributed Usability Evaluation Method (DUEM). A brief description of the method has been provided with two examples to demonstrate its use and benefits.

REFERENCES

ACM SIGCHI (1992) ***Curricula for Human-Computer Interaction***, ACM Press**.**

Bailey, R.W. (1993) Performance vs preference, in: *Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting*, pp. 282-286.

Beyer, H. and Holtzblatt, K. (1998) ***Contextual Design: Defining Customer-Oriented Systems*, Morgan Kaufmann Publishers.**

Beyer, H. and Holtzblatt, K. (1999) Contextual Design, **interactions**. 6, pp. 32-42.

Bødker, S., 1991, **Through the Interface: A Human Activity Approach to User Interface Design**, Lawrence Erlbaum.

Borgholm, T. and Madsen, K.H. (1999) Cooperative Usability Practices, **Communications of the ACM**. 42, pp. 91-97.

Dumas, J. S. and Redish, J. C. (1993) *A Practical Guide to Usability Testing*, Ablex Publishing.

Engeström, Y. and Middleton, D. (Eds), 1996, **Cognition and Communication at Work,** Cambridge University Press.

Engeström, Y. (1987) **Learning by Expanding: An activity-theoretical approach to developmental research,** Orienta-Konsultit.

Engeström, Y. (1995) Polycontextuality and Boundary Crossing in Expert Cognition: Learning and Problem Solving in Complex Work Activities, **Learning and Instruction**. 5, pp. 319-336.

Gould, J. D. and Lewis, C. (1985) Designing for Usability: Key Principles and What Designers Think, **Communications of the ACM.** 28**:**3, pp. 300-311.

Hartson, H. R., Andre, T. S. and Williges, R. C. (2001) Criteria For Evaluating Usability Evaluation Methods, **International Journal of Human-Compute Interaction**, 13:4, pp. 373-410.

Hertzum, M. and Jacobsen, N.E. (2001) The Evaluator Effect : A Chilling Fact About Usability Evaluation Methods, **International Journal of Human-Computer Interaction**. 13:4, pp. 421-444.

Holleran, P.A. (1991) A methodological note on pitfalls in usability testing, **Behaviour & Information Technology.** 10:5, pp. 345-357.

Hutchins, E. (1995) **Cognition in the Wild**, MIT Press.

Jacobsen, N. E., Hertzum, M. and John, B.E. (1998) The evaluator effect in usability tests, **CHI 98 Conference Summary,** ACM Press, pp. 255-256.

Jeffries, R., Miller, J. R., Wharton, C. and Uyeda, K. M. (1991). User interface evaluation in the real world: A comparison of four techniques, **Proceedings of CHI 91**, ACM Press, pp. 119-124.

Kaptelinin, V. (1994) Activity Theory: Implications For Human Computer Interaction, in: **Human-Machine Communication For Educational Systems Design**, M.D. Brouwer-Janse and T.L. Harrington, eds., Springer-Verlag.

Karat, J. (1997) User-centred software evaluation methodologies, in: **Handbook of Human-Computer Interaction,** Helander, M. G., Landauer, T. K. and Prabhu, P. V., eds., Elsevier Science, pp. 689-704.

Kling, R. and Iacono, S. (1989) The Institutional Character of Computerized Information Systems, **Office: Technology & People**. 5:1, pp. 7-28.

Kuutti, K. (1996) Activity Theory as a Potential Framework for Human-Computer Interaction, in: **Context and Consciousness: Activity Theory and Human Computer Interaction**, B. Nardi, ed., MIT Press.

Leontiev, A. N. (1981) **Problems of The Development of The Mind**, Progress Publishers.

Lewis, J. R. (2001) Current Issues in Usability Evaluation, **International Journal of Human-Computer Interaction**. 13:4, pp. 343-350.

Mayhew, D. J. (1999). **The usability engineering lifecycle: A practitioner's handbook for user interface design**, Morgan Kaufmann Publishers.

Molich, R., Thomsen, A. D., Karyukina, B., Schmidt, L., Ede, M., van Oel, W. and Arcuri, M. (1999) Comparative Evaluation of Usability Tests, in: **Proceedings of CHI'99**, ACM Press.

Molich, R., Bevan, N., Curson, I., Butler, S., Kindlund, E., Miller, D. And Kirakowski, J. (1998) Comparative Evaluation of Usability Tests, in: **Proceedings of the UPA Conference**, pp. 189-200.

Nardi, B. (1996) Activity Theory and Human-Computer Interaction, in: **Context and Consciousness: Activity Theory and Human Computer Interaction,** B. Nardi, ed., MIT Press.

Nardi, B. A. and O'Day, V. L. (1999) **Information Ecologies: Using Technology with Heart**, MIT Press: Cambridge.

Nielsen, J. (2000) Why you only need to test with 5 users, **Alertbox**. March 19.

Nielsen, J. (1993) **Usability Engineering**, Academic Press.

Nielsen, J. (1990) Designing for International Use, in: **Human Factors in Computing Systems, Proceedings of CHI'90,** ACM Press, pp. 291-294.

Nielsen, J. and Landauer, T. (1993). A mathematical model of the finding of usability problems. **Proceedings of INTERCHI '93**, ACM Press.

Nielsen, J. and Phillips, V. L. (1993). Estimating the relative usability of two interfaces: Heuristic, formal, and empirical methods compared, **Proceedings of INTERCHI 93,** ACM Press, pp. 214-221.

Nielsen, J. and Molich, R. (1990) Heuristic Evaluation of User Interfaces, in: **Proceedings of CHI'90**, ACM Press, pp. 249-256.

Norman, D. A. (1988) **The Psychology of Everyday Things**, Basic Books.

Polson, P. G., Lewis, C., Rieman, J. and Wharton, C. (1992) Cognitive Walkthroughs: A method for theory based evaluation of user interfaces, **International Journal of Man-Machine Studies**. 36, pp. 741-773.

Rubin, J. (1994) **Handbook of usability testing: How to plan, design, and conduct effective tests**, Wiley.

Rhodes, J. S. (2000) **Trouble in Paradise: Problems Facing the Usability Community**, (April 17, 2003); http://webword.com/moving/death/ html.

Shackel, B. (1986) Ergonomics in Design for Usability, in: **People and Computers: Designing for Usability, Proceedings of the 2nd Conference of the British Computer Society Human Computer Interaction Specialist Group**, Harrison, M. D. and Monk, A. F., eds., Cambridge University Press, pp. 45-64.

SINTEF Group (2002) The Foundation for Scientific and Industrial Research at the Norwegian Institute of Technology, (March 15, 2003); http://www.oslo.sintef.no/avd/ 32/3270/brosjyrer/ engelsk/6.html.

Spinuzzi, C. (1999) Grappling with distributed usability: A cultural-historical examination of documentation genres over four decades, in: **Proceedings of the 17th annual international conference on Computer documentation**, New Orleans, pp. 16-21.

Spool, J. and Schroeder, W. (2001) Testing Websites: Five Users is Nowhere Near Enough, in: **Proceedings of CHI'2001**, ACM Press, pp. 285-286.

Sweeney, M., Maguire, M. and Shackel, B. (1993) Evaluating user-computer interaction: A framework, **International Journal of Man-Machine Studies**. 38, pp. 689-711.

Thimbleby, H. (1990) *User Interface Design*, Addison Wesley, Harlow (UK).

Thomas, J.C. and Kellogg, W.A. (1989) Minimizing Ecological Gaps in Interface Design, *IEEE Software*, 6:1, pp. 78-86.

Thomas, P. and Macredie, R. D. (2002) Introduction to The New Usability, **ACM Transactions on Computer-Human Interaction**, 9:2, pp. 69-73.

Tullis, T. S. (1993) Is user interface design just common sense?, in: **Proceedings of the 5th International Conference on Human- Computer Interaction**, pp. 9-14.

Virzi, R.A. (1992) Refining the Test Phase of Usability Evaluation: How Man Subjects Is Enough?, **Human Factors**, **34**:4, pp. 457-468.

Vygotsky, L. S. (1978) **Mind in Society**, Harvard University Press.

Whiteside, J., Bennett, J. and Holtzblatt, K. (1988) Usability engineering: our experience and evolution, in **Handbook of Human-Computer Interaction**, M. Helander, ed., North-Holland.

Wixon, D. and Wilson, C. (1997) The usability engineering framework for product design and evaluation. In Helander, M. G., Landauer, T. K. and Prabhu, P. V. (eds) **Handbook of Human-Computer Interaction,** 2nd edition, Elsevier Science, pp 653-688.