Artifacts, Tools and Generalizing Usability Test Results

Pekka Reijonen and Kimmo Tarkkanen

Information Systems Science, Turku School of Economics, University of Turku,
Rehtorinpellonkatu 3, 20500 Turku, Finland
kimmo.tarkkanen@utu.fi

Abstract. Usability testing has gained a rather stable status as a method for usability evaluation even though it has both low reliability and validity. The sources of result variance are well acknowledged among researchers and practitioners. However, the validity problem has not been explicated or exemplified although it is frequently discussed in the literature how the results of usability tests should be interpreted and to what extent results are generalizable. We employ Activity Theory and a case example to argue that the validity problem is mainly caused by the fact that what we are testing are artifacts and what people are using in their real life activities are tools and these two entities are qualitatively different. Basing on our analysis, the effects of the reliability and validity problems on the application of usability testing and its role as one of the tools in the design process are discussed.

Keywords: Usability testing, validity, generalization, reliability, activity theory

1 Introduction

Reliability and validity are the measures used in ascertaining the quality of evaluation instruments. Reliability is used in assessing if the measuring tool produces consistent results and validity in ascertaining if the tool is measuring what it is supposed to measure. In the past 20+ years, usability testing has gained popularity in such a manner that some kind of usability testing or evaluation plays some role in practically all software development projects. Usability tests are carried out by different actors during different stages of product's lifecycle with techniques ranging from a heuristic evaluation to laboratory and field tests [1]. The background knowledge of the testers varies considerably from layman to usability experts [2] and even automated asynchronous usability tests have been introduced [3]. When the diversity of usability testing procedures and actors are combined with the fact that there exists also many somewhat different, although overlapping, definitions of the usability concept [4], it is not always clear, how the results of a usability test should be interpreted [5] and to what extent these results are generalizable [6]. Repeatedly, empirical studies wonder why the usability of the system differs in pre- and post-implementation phases despite extensive and varied empirical usability and user research efforts (cf. [6]).

In this paper, we attempt to shed light on the interpretation and generalization issues of usability test results. With the activity theoretical support [e.g. 7] we explicate why and how generalizing the results of usability testing to real life situations is not a straight forward procedure. Our exploration is conceptual although as a case example, we refer to the famous empirical study by Suchman [8]. The rest of the paper is organized as follows. In the following chapter the basic variables of usability testing are listed and their effect on reliability is shortly discussed. The next chapter is dedicated to the validity issues, i.e. what type of generalizations are made from the usability test results and how justifiable these generalizations are. In the last chapter, the remarks and arguments about the reliability and validity issues are discussed and a constructive way for using usability test results is outlined.

2 Variables of Usability Testing

According to their use, usability tests can be divided into two broad categories, formative and summative evaluation. Summative evaluation is used, for example, to verify that the delivered product fulfills the usability criteria or to compare the usability of two or more products. The aim of a formative evaluation is during the development process to ascertain that the brewing artifact will meet the predefined usability criteria, i.e. to enhance the usability of the final artifact by helping to remove usability problems. Quite independently of the usability method used or the purpose of the evaluation, the primary output of a usability test is a list of usability problems that form the basis for recommended changes in design [5], [9]. Other often observed dependent variables of usability tests include, for example, time and subjective satisfaction [1]. A usability problem is the most often observed variable in usability tests [10], [5], [2]. In practice, usability problems are identified through the direct observation of users' verbal and non-verbal behavior or indirectly by the evaluator [11]. The trouble with the concept 'usability problem' is that it is elusive and therefore it has no generally accepted definition, i.e. every test administrator seems to use their own criteria. If the definition of the usability problem is vague, all the comparisons based on the number or quality of problems are also vague. This alone partially explains the confusing results attained when different test groups have evaluated the same artifact [12], [9]. Even the broadest possible definition of usability problem, "anything that impacts ease of use - from a core dump to a misspelled word" [13, p.121], leaves the responsibility to the evaluator. In other words, it does not remove the basic cause of confusion, namely that it is up to the evaluator to decide, what impacts ease of use or hinders use in general [14]. Execution time can be reliably and consistently measured, yet it is less interesting variable, as formative usability tests are mainly used for diagnostic purposes during the development process and aim at design changes. User satisfaction, gathered in interviews and standardized questionnaires, is important not only in detecting usability problems, but specifically in interpreting the causes of the problems. However, data gathering methods are mostly used rather informally and inconsistently as explicated in the analysis of the thinking aloud method by [15].

The tested artifact is the main independent variable in a usability test as all the variation observed in the dependent variables is supposed to be caused by the attributes of the artifact [16,17,18]. The attributes of the artifact are not, however, the only independent variable causing variation in the dependent variables. Although the artifact is kept constant, different subjects detect and experience different usability problems [9,10]. Interpretation of test results becomes additionally trickier when two remaining independent variables, test task and test arrangement are also taken into account. At the minimum, the attributes of the test task include the number of tasks (count), type, and coverage, which affect the number and quality of problems found [19]. Despite the fact that there exist numerous general recommendations [1] on how to run usability tests, test arrangements and procedures are far from standardized. Attributes like administrator, testing premises (laboratory, field), observation method (think aloud, observation, video recording), and test situation (pairwise or single subjects or a group), and training before the test session are sources of variability and low reliability (see [20,21]).

It must be kept in mind that a typical usability test lacks control group so it does not qualify even as the simplest possible experimental design [22]. In practice this means that one must be very careful when interpreting any causal relations between the variables.

3 Generalization of Test Results

As shortly discussed above, there are many variables that can cause uncontrolled changes in the output of the test, i.e. lower its reliability and in principle unreliable results should not be generalized at all [22]. Usability tests are, however, reliable in one respect; they all consistently produce a list of problems with accompanying recommendations. Generalizations are based on the expectation that the external validity of the study is high, i.e. the results hold for other test situations, subjects, times, and environments [23]. Four types of generalizations are routinely done from usability test results (Table 1). The first one is the generalization from the used test tasks to all possible test tasks. For several reasons, often economical, only the parts of software that are considered the most important by some influential actor or the test administrator are tested. This selection is often based on the estimated usage frequency, i.e. the most used software features are tested. It is certainly important that the most frequently used features can be used fluently. The paradox here is that during use users get lots of training in the most used features, but the least used features can cause problems later as they are never properly learned because of infrequent use. For this reason precisely the infrequently used but important features should be easy to learn or rather self-explanatory. The other problem with the generalization from the used test tasks to all the possible test tasks is the type of tasks, especially their breadth (from simple tasks with definite answers to more general ones, see e.g. [24]). For example, in the case of simple tasks with definite solutions the subjects do not need to understand the task flow or how a work process is carried out with the artifact. This problem is confounded with the generalization from the artifact test to the tool use and will be discussed farther down.

Even though the effect of the subjects might be less central to the results of usability testing than to those of user experience [25] the representativeness of the test subjects must be considered in every usability test. It is commonly [1] recommended that the test subjects should be selected from the future users. In general, by following this principle, it is rather safe to generalize the test results to the whole user population. The recommendation is based on the assumption that the target user group is homogenous, i.e. all users have about equal IT skills and interpret the work processes in a similar manner. The situation is not, however, always this straight forward. For example, the new artifact can be designed to help in changing the old work process and does not support the existing one, hence the knowledge of the old work process can actually be a hindrance instead of an advantage. For example, home care nurses were puzzled when they did not find a similar detailed list of work tasks on a mobile device as they were used to get in print from the old desktop system [24]. In the study reported by Suchman [8], the new mechanical parts of the photocopier and the change in the way it handled the originals while making double sided copies were one of the main causes of confusion. In other words, a new version of an artifact can be easier to use for a total novice than for a novice who knows the earlier versions of the artifact or the existing work procedures.

Table 1. Generalizations of usability test results.

Test attribute	Generalization to
Used test tasks (partial test, task coverage)	All possible test tasks or the whole software
Test subject	All users
One test arrangement	All test arrangements
Artifact test	Tool use

Test arrangements, i.e. how the test is actually carried out, can vary substantially and hence can have an effect on the results. In their explorative field research Boren and Ramey [15] studied thinking aloud method by observing seven usability experts in two professional organizations and found that there is considerable variation in how the method is applied even in the same organization. For example, there were variations in how the participants (subjects) were instructed to think aloud, how and in what pace reminders were given, how practitioners intervened, and how verbal test protocols were treated in interpreting the results. This study clearly shows that the widely applied thinking aloud method is used inconsistently and there are differences between practice and theory, but it does not give any hint of the consequences of the differences between the test results. The data collected from research literature by [20] also reveal that there is a considerable evaluator effect in usability testing, i.e. irrespective of the usability evaluation method (cognitive walkthrough, heuristic evaluation, thinking aloud) different evaluators report substantially different sets of usability problems and seem to rank the severity of the problems differently. Basing on the research on usability practice, [26] conclude that usability evaluation inevitably includes a lot of value judgments and hence experience and competence of usability practitioners is crucial and "Regardless of qualifications, success in systems development indicates a high level of intelligence" [26, p.961].

The most extensive series of studies comparing the usability test results of different test administrators has been carried out by Rolf Molich [9]. The goal of these comparative usability evaluations (CUE) has been, among other things, to find out to what extent the usability evaluation results are reproducible. The number of test teams has varied from four (CUE-1) to seventeen (CUE-4) and the number of usability issues reported only by single teams has varied from 95 percent (128 of 141 issues) to 60 percent (205 of 340 issues). For example, in the CUE-4 study, none of the issues were mentioned by all teams and 6 of the 340 issues (1.8%) were reported by the half of the teams (8). In other words, the results of usability evaluations were far from reproducible even though the test teams received the same client scenario describing the main goals of the usability evaluation. The teams were allowed to use their preferred test method so the arrangements varied in many respects, like the number of test sessions, number and type of tasks and scenarios, and the testing premises. The focus of the research was, however, on the results of the tests rather the methods, so it is impossible to infer the effects of the different independent variables from the data.

The most significant, and maybe the least considered, generalization of usability tests is that the results attained in the test are more or less directly applicable to real use situation. This generalization is made commonly implicitly as its rationale lies in the center of the whole idea of usability testing: in (formative) usability testing, usability problems are detected and when these problems are removed, the artifact is more usable. This is admittedly true, if all the other independent variables except the artifact are kept constant when retesting the artifact. In other words, in the retest the subjects, the tasks, and the test arrangements and procedures are the same as in the original test. If any of these independent variables is changed, we do not know any more if the observed changes are caused by the changes in the artifact or the changed independent variable. We maintain that the change of the arrangement, from testing an artifact to using a tool, is so drastic that generalization should be made very carefully. By tool, it is meant “something (as an instrument or apparatus) used in performing an operation or necessary in the practice of a vocation or profession” [27].

Nielsen [17] made in his early usability definition a clear distinction between usability and utility which are the constituents of usefulness, i.e. “whether the system can be used to achieve some desired goal” [17, p.24]. According to this definition, utility concerns the functionality of the system and usability is the question about how well users can use this functionality. The distinction between utility and usability is not always straightforward and the examples of the benefits of usability engineering given by Nielsen [17, p.2] point actually more to utility than usability. For example, the damage claim system of an insurance company was designed so that the whole transaction should be carried out completely and if interrupted, the transaction must be started from the very beginning and all previously input data was lost. This caused considerable trouble in the offices and required workarounds, but it is a question about the functionality of the system, not its interface. This example also uncovers two other issues that have to do with the generalization of usability test results to work practices. The first is that the results were not obtained in a usability test, but by observing the use situation and interviewing the users. The second is that the users were not novices as in a typical usability test, but had been using the system long enough to create workarounds for managing the shortcomings of the system. In other words, the observations do not come from artifact testing, as in traditional usability testing, but from tool use in work practices.

4 A Case Example

We explicate the difference between testing an artifact and using a tool by referring to probably the most ever cited single usability test, namely the research carried out by Suchman [8], [28], [29]. This study has not been called a usability test by Suchman, or to our knowledge by anybody else neither, but its empirical part is anyhow a usability test carried out in a laboratory. Actually, 'usability' was not even mentioned, at least not in the original report from the year 1985 [8], and there are at least two reasons for that. First, the goal of the study was not to detect usability problems and hence improve the usability of the artifact but to better understand human-machine communication. Second, at the beginning of the 1980s usability or usability testing had not received the kind of attention as they have today. In fact, this study is one of the first ones that draw attention to the problems the users 'in the field' had when applying computer based artifacts in their work. In that time, actually, "whitewater canoeing" [29, p.19] was neither called whitewater canoeing, but "to run a series of rapids in a canoe" [28, p.52].

The actual research was carried out in a laboratory, where a video camera was set up to record the interaction of test subjects with the photocopying machine. All the subjects were novices, i.e. they did not have received training in the use of the machine nor knew its somewhat different functionality in making two sided copies of a bound document. The subjects were given the test task (e.g. make two-sided copies of a bound document) and then left alone to work with the machine. The discussion protocols were transcribed from videotapes and analyzed. When the data is considered as a usability test protocol, it is obvious that the majority of the problems was caused by the fact that the test subjects were novices and could not even recognize the parts of the machine nor know its functions. For example, the subjects did not

- know what Bound Document Aid (BDA) is [8, p.91 and p.111] and if "the latch labeled Bound Document Aid" should be pulled or pushed [8, p.96] or what is the document cover [8, p.92 and p.111]
- find the start button [8, p.104 and p.116]
- know that contrary to the older machines, in this machine all pages of a multiple pages unbound document must be loaded at once and not one-at-a-time [8, p.117]

The study clearly shows that the copying machine was not self-explanatory for a novice user and the evidence from Suchman's pilot studies also throw light on the concept 'novice'. In a video recording, two men try to make two-sided copies of a research to their colleagues and their behavior looks more like a deliberately comic performance than work practice [30]. The men in the video clip were the senior computational linguist at PARC, Ron Kaplan, and one of the founders of the AI (Artificial Intelligence) movement, Allen Newell. In other words, in front of new, computer based equipment nearby everybody is a novice. This example, as well as Suchman's [8] other empirical data, clearly demonstrates the difference between an artifact and a tool: the subjects are obviously trying to make sense of an artifact and not performing a routine work task using a tool. This interpretation is strengthened by the fact that this kind of behavior, i.e. keep trying to make copies for hours, can usually take place only in an experimental setting. In a work setting, i.e. at a quite normal workplace, the workers would have kept trying for a few minutes and then asked for help from their co-workers, help-desk, or invented a functioning workaround.

The difference between testing an artifact and working with a tool can be made explicit by identifying the goal or motive and the actions of the observed behavior. We utilize Activity Theory in clarifying this point as it offers suitable concepts and structures for describing human goal oriented behavior, for example when a human is carrying out work tasks in a specified context using appropriate means like artifacts and/or tools [7], [31]. The basic unit of analysis is an activity that includes a minimal meaningful context for actions, is directed towards an object and turning the object into an outcome is the motive of the activity. The basic activity system consists of a subject (actor), an object, and tools. The actor is not manipulating the object directly, but doing is mediated by the tool (artifact) that is the result of a historical development and also sets limits for doing. An activity is realized by a series of actions carried out by the actor. Every action has a specific goal and the subject is aware of the goal she wants to achieve. Depending on the situation, the same activity can be realized by different actions and the same action can be part of different activities. Through human learning, an action can, and usually will, collapse into an operation, which is a habitual routine that needs less conscious attention than an action and is adjusted to the specific conditions. When an action turns into a routine operation a new, broader action is formed and it includes the operation as a subpart. If the conditions change, for example in the case of a breakdown, the subject can return the operation back to the conscious action, in other words, an operation is not a conditional reflex.

Table 2. An activity theoretical interpretation of two different behaviors: testing an artifact and working with a tool. Data of Suchman [8], [28] re-interpreted by the authors.

Activity system	**Testing an artifact**	**Working with a tool**
Subject	Test subject, knows how to behave in a test situation, may know the work practices, tool is typically novel	Worker, knows how a tool is used in work practices
Object (Motive) → Outcome	Use a copying machine → A pile of double-sided copies	Share knowledge → Information delivered to colleagues
Tool	Copying machine	Copied research paper
Level of activity		
Activity	Taking two-sided copies of a bounded research paper	Taking part in a collaborative research endeavor
Action	Reading instructions and trying to make sense of the interface	Taking two-sided copies of a bounded research paper
Operation	Identifying controls and parts, pressing buttons	Making appropriate operations in appropriate order

When analyzing the activity system, it can be observed that the subjects in the test situation are test subjects who do not know how the artifact is used, but can be aware of the work practices (Table 2). When a tool is used in a work practice, the subject is a worker who knows how the tool is applied in performing the work practices. The object of the activity in the test situation is the use of the copying machine and the outcome is a pile of double-sided copies of a bounded document created with the copying machine (tool or actually an artifact). In the work situation, the object (motive) of the activity is knowledge sharing and the outcome is information delivered to colleagues. The tool used for knowledge sharing is the copied research paper, not the copying machine.

As the activity systems in these two situations are different also the deeds are different on the different levels of the activity. In the test situation, the activity is solely the making of copies, whereas in the work situation the activity is a collaborative research and copy-taking is just one action in this activity. The actions in the test situation taken by the subject concentrate in sense making, i.e. the subject reads instructions and proceeds step-wise in a trial and error mode. When a worker uses the copying machine as a tool, the action is simply taking two-sided copies of a bounded research paper. The operations used in the actions of the test situation consist of reading instructions in order to identify parts and find appropriate controls and buttons to press. In the work situation, the worker takes copies as usual, i.e. by making appropriate operations in appropriate order and maybe talking on the phone simultaneously. According to our interpretation these two situations, artifact test and tool use, are so profoundly different that the results of the test cannot as such be validly generalized to the real work environment. This issue is further elaborated in the last chapter.

5 Discussion

Usability testing has obtained a rather stable status among the methods applied in systems design and development process. The rationale and justification of testing is rather straight forward: when the flaws of the design detected in usability testing are removed the designed artifact suits better its purpose. This is a logically sound conclusion and on some level also an appropriate interpretation but at a closer look there are several factors that must be taken into account when interpreting the results of a usability test and generalize the results to other environments than the test situation.

To begin with, considerable variation in the results of usability tests is introduced by the variations in the independent variables, i.e. artifact, subject, task, and test arrangements and procedures. As the CUE-series of experiment [9] show keeping two variables, the artifact and testing methodology, constant, does not much reduce the variability of the results. This is understandable as the two other independent variables can vary freely. For example, the subjects’ skills and knowledge may be very different and the tasks can vary from simple small tasks to longer work processes.

The typical dependent variables, i.e. the measured or observed things, are usability problems, execution time, and different subjective measures like satisfaction. From these variables, only execution time can be reliably and repeatedly measured and compared with the presupposition that the independent variables, at least tasks and test procedures, are kept constant. There have been efforts to define the concept of usability problem more precisely [18] and systemize the extraction of usability problems from test data [32] but their effects on practices have been meager. The same holds for test procedures, as there seems to be differences, for example, in the use of think aloud method even inside the same organization [15].

As the analysis of the usability test variables show (Chapter 2), they are all potential sources of variation in the results of a usability test, thus the variability of results found in the CUE studies [9] is understandable. Despite the fact that different usability tests produce different recommendations, the recommendations are meant to be applied in the design process in order to improve the usability of the tested artifact. It is further expected that following the recommendations has some positive effect on the actual use situation of the tested artifact [6]. This generalization means that the test tasks, subjects, and arrangements are taken as a representative sample of their respective universe. If the subjects are selected carefully, the test tasks are formed sensibly, and the tests are carried out following the recommended procedures, the generalization can be justified. There is, however, one generalization which is more questionable: how the results from testing an artifact can be generalized to the use of a tool.

As pointed out earlier (see Table 2), there is a qualitative difference between the test situation and the use situation. This difference is based on the difference between an artifact and a tool. According to the dictionary definition, an artifact is "something created by humans usually for a practical purpose" and a tool is "something (as an instrument or apparatus) used in performing an operation or necessary in the practice of a vocation or profession" [27]. This difference was insightfully described by Butler in the late 1800-hundreds [33]: "Strictly speaking, nothing is a tool unless during actual use. Nevertheless, if a thing has been made for the express purpose of being used as a tool it is commonly called a tool, whether it is in actual use or no. We see, therefore, matter alternating between a toolish or organic state and an untoolish or inorganic. Where there is intention it is organic, where there is no intention it is inorganic." According to our interpretation, it is exactly this alternation of a product between the untoolish and toolish states that make the direct generalization of the test results to the use situation in many cases unjustified or at least somewhat difficult.

In order to clarify the problem of the contextual generalization we refer to the case study reported by Riemer and Vehring [6], where the use of an IP-based telephony system was observed and the users were interviewed in their workplace context. The aim of the study was to enhance the functionality and especially the usability of the system as the changes made according to the recommendations of a recent laboratory based usability test were not received well by the users. It turned out that in different contexts the functions of the software were utilized in a varying degree and hence also the hardware varied from conventional phones to wireless headsets. As there was no single unified use context or hardware configuration the authors conclude that in this case "establishing a notion of usability as a characteristic of the software turned out an impossible task" [6, p.7]. Based on the observations and interpretations it is further maintained that laboratory based usability test can be "counterproductive, as it might produce results that are detrimental to the ways in which usability manifests in the sociomaterial use context" and "usability should be treated as a distinctly contextual phenomenon" [6, p.13]. This study clearly explicates the problem of contextual generalization, but the offered solution, the development of contextual usability testing methods, also has several drawbacks.

A widely accepted standard defines usability as "the 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" [34]. If this definition is taken literally, it means that the results from a usability test are valid only in the testing environment (specified context of use) and not in the actual use environment regardless of how well the subjects or test tasks represent the actual use situation. In this case, the validity and generalizability of the usability test results would be the highest in the situation where the artifact never achieves the status of the tool. This happens when users 1) do not learn to use the artifact properly and 2) never integrate its use into their work practices. This may happen in situations where the artifact is used seldom, like connecting a laptop to the presentation equipment of a lecture room in a strange environment or using a web shop or other web application for the first and only time.

Contextual usability testing would mean that the usability of the same artifact should be tested separately in every sociomaterial use context, because "usability manifests as an aspect of this sociomaterial use context, with which software and hardware become entangled" [6, p.2]. This would also mean that the UCD-type of software design and development is impossible as usability can be assessed only after the software has been implemented into and is used in a certain environment. In other words, this would be a move backwards to the waterfall model of systems development.

Carrying out usability tests on the field instead of a laboratory inevitably introduces some additional independent variables that cannot be controlled or even reliably measured. For example, users' skills and knowledge are at different level, the system has been implemented into the work processes in a certain way, division of labor varies according to the organization, workarounds have been formed to overcome obstacles, etc. In other words, the artifact-tool problem can not be generally solved simply by changing the test premises from a laboratory to a work place as an artifact does not turn into a tool when it is placed in a different location but through human learning when the artifact is used in work practice (cf. [35,36,37]). Testing an artifact or a tool in a certain sociomaterial context would surely produce results that are applicable in that specific environment, but these results are of less value for the artifact vendor as it is 1) impossible to know to what other environments these results apply, 2) laborious to produce and maintain different artifacts for every possible application environment, and 3) expensive and technically difficult to implement changes in the artifact that is already ready and in use.

If the focus of usability testing is changed from the attributes of the artifact to the utility of the artifact's functions in a certain environment, the question is, is that usability testing or something else. Nonetheless, the usability tests of artifacts can still have relevance in the design process and for the usability of the final product. This standpoint presupposes that artifacts have attributes that can be evaluated independently and without a direct connection to all aspects of its future use environment or its future utility in that environment. This kind of usability testing has been practiced for over 20 years and we maintain that most of the results have been more useful than harmful for systems design. For example, with the help of usability testing it is possible to lower the skills and knowledge requirements of the users by changing the attributes of the artifact in a more comprehensive direction, i.e. less resource is required for training the users irrespective of the use environment (learnability, memorability, errors). The same argument holds for the task execution time, i.e. in most environments, it is beneficial to use less time for a given task (efficiency). Similarly, we maintain that observing tool use in a specific context after the product launch has a firm place in the overall product lifecycle. Thus, artifact testing and tool use are not in competition, but should be acknowledged in the development lifecycle as qualitatively different means to obtain better usability.

It is a reasonable requirement that the procedures of testing and the ways conclusions are drawn are made explicit and methods are applied consistently, but this does not make usability testing an exact science as claimed by Nielsen [17, pp.26-27]: "Only by defining the abstract concept of "usability" in terms of these more precise and measurable components can we arrive at an engineering discipline where usability is not just argued about but is systematically approached, improved, and evaluated (possibly measured)." This approach would possibly increase the reliability of usability testing, but would not solve the problem of validity, i.e. "whether the usability test in fact measures something of relevance to usability of real products in real use outside the laboratory" [17, p.169]. An easy way to guarantee the validity of a tool is to define the measured construct through the measuring tool as made by Nielsen [17, p.23]: "I tend to use the term "usability" to denote the considerations that can be addressed by the methods covered in this book". Unfortunately, this is actually about the level the validity of usability testing has been evaluated. The main reason for this is the high face validity of the usability testing methods. In other words, both laymen and most experts agree upon the fact that the tests measure exactly the right concept [38].

As we have explicated earlier, the generalization from artifact testing to tool use is not a straight forward procedure, but if usability tests are planned carefully, the tool use situation can be to a certain extent simulated by testing an artifact. One possibility is to use open ended tasks that force the subjects to create smaller tasks on the fly in the test situation while simulating their work practices [24]. Another approach with at least some face validity is a procedure where the business goals of the system are considered explicitly in planning and reporting usability tests [39]. Direct and reliable evidence of the effects of artifact test findings on real work situations is hard to attain, but it is, however, possible to integrate usability testing into the redesign of an existing system as done by [40]. The rationale is that usability problems should not be addressed in isolation but integrated into the redesign process as one of the sources producing design alternatives. This is extremely important as in order to be beneficial at all the results of the usability tests should be taken into account, which does not seem always happen. For example, [9] noted that two years after the comprehensive series of usability tests of a web site only 4 of the 26 key problem issues had been apparently solved, even though there had been resources for the development as some new features had been added. As a remedy, [9] suggests that the evaluator should interact closely with the designers and not just deliver a test report. This also implies a hint of the way the scope of usability testing should actually be defined, i.e. it should shift from the number of problems to the effects on the design, or as proposed by [41, p.105], "the true utility of methods lies in their ability to influence the design of the application being evaluated." In other words, usability testing should be comprehended as an inherent part of design that helps to create design alternatives for an artifact and not as an exact method for enhancing the utility or even the effectiveness of a tool. This definition would also be more realistic and easier to verify than the prevailing one, i.e. the results of an artifact test can be generalized to a tool use situation.

In this theoretically oriented paper, we have used second hand empirical data to highlight the qualitative difference between an artifact and a tool. To our knowledge, this difference has not been considered in usability testing literature and in activity theoretical literature these concepts have been applied interchangeably. If this difference is considered as a new independent variable in usability testing, it can to a great extent explain the different results obtained when the same computer based system has been tested in the laboratory (using an artifact to carry out test tasks) and observed in the field (using a tool in routine work tasks). The main problem with the proposed distinction is how we in practice know if something is an artifact or a tool for its user. One way to determine this is to use an activity theoretical approach: when the user carries out conscious actions she is using an artifact and when the actions have collapsed into routine operations she is using a tool, in other words, humans are using tools and not artifacts in their routine work tasks. Basing on this, we can rather safely state that a computer based system is an artifact when an actor uses the system for the first time, like the subject in a conventional usability test. Unfortunately, it is much more difficult to empirically ascertain when an artifact has become a tool for the user. This is one of the questions that should be clarified in the future research.

References

1. Barnum, C.M.: Usability Testing Essentials: Ready, Set… Test! Morgan Kaufmann, Burlington (2011)
2. Hvannberg, E.T., Law, E.L-C., Lárusdóttir. M.K.: Heuristic Evaluation: Comparing Ways of Finding and Reporting Usability Problems. Interacting with Computers 19, 225–240 (2007)
3. Andreasen, M.S., Nielsen, H.V., Schrøder, S.O., Stage, J.: What Happened to Remote Usability Testing? An Empirical Study of Three Methods. Proc. CHI 2007, pp. 1405-1414. ACM Press (2007)
4. Alonso-Ríos, D., Vázquez-García, A., Mosqueira-Rey, E., Moret-Bonillo, V.: Usability: A Critical Analysis and A Taxonomy. International Journal of Human-Computer Interaction 26, 1, 53-74 (2009)
5. Hornbæk, K.: Usability Evaluation as Idea Generation. In: Cockton, G.G., Hvannberg, E.T., Law, E. (eds.) Maturing Usability: Quality in Software, Interaction and Value. pp. 267-286. Springer (2008)
6. Riemer, K., Vehring, N.: It's Not a Property! Exploring the Sociomateriality of Software Usability. In: Proceedings of the International Conference on Information Systems (ICIS), Phoenix, Arizona, pp. 1-19. (2010)
7. Kuutti, K.: Activity Theory as a Potential Framework for Human-Computer Interaction Research. In: Bonnie A. Nardi (Ed.) Context and Consciousness, pp. 17-44. MIT Press (1995)

8. Suchman, L.A.: Plans and Situated Actions. The Problem of Human-Machine Communication (Thesis). XEROX PARC. ISL-6. (1985)
9. Molich, R., Dumas, J.S.: Comparative Usability Evaluation (CUE-4). Behaviour & Information Technology 27, 3, 263-281. (2008)
10. Nielsen, J., Landauer, T.K.: A Mathematical Model of the Finding of Usability Problems. Proc. CHI, pp. 206-213. ACM (1993)
11. Følstad, A., Law, E.L.-C., Hornbæk, K.: Analysis in Practical Usability Evaluation: A Survey Study. In: Proc. CHI 2012, pp. 2127–2136. ACM Press (2012)
12. Molich, R., Ede, M.E., Kaasgaard, K., Karyakin, B.: Comparative Usability Evaluation. Behaviour & Information Technology 23, 65-74 (2004)
13. Jeffries, R., Miller, J.R., Wharton, C., Uyeda, K.M.: User Interface Evaluation in the Real World: A Comparison of Four Techniques. Proc. CHI 1991, pp. 119-124. ACM (1991)
14. Vermeeren, A., van Kesteren, I., Bekker, M.: Managing The Evaluator Effect in User Testing. Proc. Interact 2003, pp. 647-654. IOS Press (2003)
15. Boren, T., Ramey, J.: Thinking Aloud: Reconciling Theory and Practice. IEEE Transactions on Professional Communication 43, 261-278 (2000).
16. Molich, K., Jeffries R., Dumas, J.S.: Making Usability Recommendations Useful and Usable. Journal of Usability Studies 2, 162-179 (2007)
17. Nielsen, J.: Usability Engineering. Academic Press, (1993).
18. Andre, T.S., Belz, S.M., McCrearys, F.A. Hartson, H.R.: Testing a Framework for Reliable Classification of Usability Problems. Proc. Human Factors and Ergonomics Society Annual Meeting 44, pp. 573-576. SAGE Publications (2000)
19. Lindgaard, G., Chattratichart, J.: Usability Testing: What Have We Overlooked? Proc. CHI 2007, pp. 1415- 1424. ACM (2007)
20. Hertzum, M., Molich, R., Jacobsen, N.E.: What You Get Is What You See: Revisiting the Evaluator Effect in Usability Tests, Behaviour & Information Technology, 33, 2, 144-162 (2014)
21. Duh, H.B-L., Tan, G.C.B, Chen, V.H.: Usability Evaluation for Mobile Device: A Comparison of Laboratory and Field Tests. In Proc. MobileHCI, pp. 181-186. ACM (2006)
22. Drost, E.A.: Validity and Reliability in Social Science Research. Education Research and Perspectives, 38, 105-123 (2011)
23. Trochim, W.M.: The Research Methods Knowledge Base, 2nd ed., http://www.socialresearchmethods.net/kb/ (version current as of October 20, 2006) (retrieved 20.1.2015). (2006)
24. Tarkkanen, K., Reijonen, P., Tétard, F., Harkke, V.: Back to User-Centered Usability Testing. In: Holzinger, A., Ziefle, M., Hitz, M., Debevc, M. (eds.) Human Factors in Computing and Informatics. LNCS, vol. 7946, pp. 91-106. Springer, (2013)
25. Arhippainen, L.: Studying User Experience: Issues and Problems of Mobile Services - Case ADAMOS: User Experience (Im)possible to Catch?. Acta Universitatis Ouluensis. Series A, Scientiae rerum naturalium (528) (2013).
26. Woolrych, A., Hornbæk, K., Frøkjær, E., Cockton, G.: Ingredients and Meals Rather Than Recipes: A Proposal for Research That Does Not Treat Usability Evaluation Methods as Indivisible Wholes. International Journal of Human-Computer Interaction 27, 10, 940-970 (2011)
27. Merriam-Webster Online dictionary, http://www.merriam-webster.com/dictionary/
28. Suchman, L.: Plans and Situated Actions. The Problem of Human-Machine Communication. Cambridge University Press (1987)

29. Suchman, L.: Human-Machine Reconfigurations. Plans and Situated Actions. Cambridge University Press (2007)
30. Duguid, P.: On Rereading. Suchman and Situated Action. Le Libellio d' AEGIS 8, 2 Été, 3-9 (2012)
31. Bardram, J., Doryab, A.: Activity Analysis – Applying Activity Theory to Analyze Complex Work in Hospitals. In: CSCW 2011, pp. 455-464. ACM, New York (2011)
32. Cockton, G., Lavery, D.: A Framework for Usability Problem Extraction. In: Sasse, M.A., Johnson, C.V. (eds.) Proc. Interact'99, pp. 344-352. IOS Press (1999)
33. Butler, S.: The Note-Books of Samuel Butler. Edited by Henry Festing Jones. (Retrieved 8.2.2015 from http://www.gutenberg.org/ebooks/6173). (1912)
34. ISO 9241-11:1998 Guidance on Usability. International Organization for Standardization, ISO 9241-11 (1998), http://www.iso.org (1998)
35. Kjeldskov, J., Skov, M.B., Als, B.S., Høegh, R.T.: Is It Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field. In: MobileHCI 2004, pp. 61-73. Springer, Heidelberg (2004)
36. Rogers, Y., Connelly, K., Tedesco, L., Hazlewood, W., Kurtz, A., Hall, R.E., Hall, R.E., Toscos, T.: Why It's Worth the Hassle: The Value of In-Situ Studies When Designing Ubicomp. In: Krumm, J., Abowd, G.D., Seneviratne, A., Strang, T. (eds.) UbiComp 2007. LNCS, vol. 4717, pp. 336-353. Springer, Heidelberg (2007)
37. Nielsen, C.M., Overgaard, M., Pedersen, M.B., Stage, J., Stenild, S.: It's Worth the Hassle!: The Added Value of Evaluating the Usability of Mobile Systems in the Field. In: Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles, pp. 272-280. ACM (2006)
38. Nevo, B.: Face Validity Revisited. Journal of Educational Measurement 22, 4, 287-293 (1985)
39. Hornbæk, K., Frøkjær, E.: Making Use of Business Goals in Usability Evaluation: An Experiment With Novice Evaluators. Proc. CHI 2008, pp. 903-912. ACM (2008)
40. Uldall-Espersen, T., Frøkjær, E., Hornbæk, K.: Tracing Impact in a Usability Improvement Process. Interacting with Computers 20, 1, 48—63 (2008)
41. Hornbæk, K.: Dogmas in the Assessment of Usability Evaluation Methods. Behaviour & Information Technology 29, 1, 97–111 (2010)