Testing the Unknown – Value of Usability Testing for Complex Professional Systems Development

Kimmo Tarkkanen, Ville Harkke, Pekka Reijonen

Testing the Unknown – Value of Usability Testing for Complex Professional Systems Development

Kimmo Tarkkanen, Ville Harkke, and Pekka Reijonen

Information Systems Science, University of Turku, Finland
{kimmo.tarkkanen, ville.harkke}@utu.fi

Abstract. To make an impact on the design in usability testing, the test tasks are essential ingredients for the early system development process. Complex design problems are not solved by focusing on the details of a prototype and setting the scope on what is already known by the design team. Instead, the design value of usability testing is increased by deliberately relinquishing the assumptions made and implemented into a design. In the development of complex systems, usability testing with extended scope and open-ended structure, as presented in this paper with three empirical cases, delivers not only specific knowledge about the user interactions with the system, but reveals issues that, despite rigorous user research efforts, have been overlooked in the preceding phases of system development. Therefore, we suggest applying open-ended usability test tasks for testing systems in complex settings such as in the development of health care systems.

Keywords: Usability testing; Test task; Design; Complex systems; Health care

1 Introduction

Professional systems in healthcare are designed to support work that can be described as complex problem solving [1]. Complex problem solving is characterized by the unpredictability of the process, as the path to solving any given problem may differ from the path of another. Professional autonomy and the nature of activities in the health care domain introduce many kinds of varying work practices and essential workarounds [2] This natural unpredictability and complexity of the health care domain have a profound effect on designing and testing of health care systems, making the design of an optimal professional system a wicked problem (see [3]). As solutions to wicked problems are tested in practical settings [4], the obvious path to better health care systems is user involvement and user studies.

The method collection known as usability testing has retained its popularity as one way of validating the proposed design solutions within the modern system development methodologies. Even when leaning on close communication with the customer (e.g. agile methods) there are clear benefits to be realized by conducting user research [5]. In its classical form, usability testing is focused on detecting the usability problems of the software product and recommending correspondent changes to the design.

The problem-centric approach of usability testing, the validity and the reliability of found problems as well as the value of succeeding design recommendations are all questioned in the past [6,7,8,9,10]. Major challenges have been that most of the reported usability problems only confirm earlier impressions of developers [11] i.e. developers are not very interested in usability problems, nor do they react on these [9].

Therefore, the modern formative usability testing aims to influence the design process and the designed artifact with more cooperative manner with the development team than before [6,7,8]. However, put into the context of complex problem domains, the scope and focus of usability tests are too often traditional and narrow, not aiming at reviewing users' actual work in these contexts [12]. In a complex domain, software developers should search multiple and alternative contexts of use [1], explore the right design direction by generating and testing ideas, instead of trying to get the first design right [13]. Thus, to have an impact on design, usability tests will have to mirror the complexity of the problem domain in the planning and aim at revealing issues that bring developers closer to a solution to the wicked problem. In practice, this means questioning all that is known in the design – testing the unknown – where the focus is set on acquiring user knowledge for the development with scope that covers not only the design artefact, but the whole spheres of contexts of use and beyond.

In this paper, we introduce three usability tests conducted in the early development phases of new health care systems. Usability test tasks given to users were open-ended in order to broaden the scope and focus of tests on acquiring knowledge of contexts that would serve the design in the complex domain as well as possible. The cases introduce results and design issues that are beyond traditional. First, the results corrected designers' wrong assumptions about the current work practices at home care. Second, the results identified the adjunct roles of users in occupational health care that refined the whole scope of the development project. Third, the results manifested the low power of end-users and software developers in the development of the national health archive. After discussing the findings in relation to usability testing procedures and objectives in systems development, we conclude that usability tests can influence complex systems development in multiple and unexpected ways. We maintain that in supporting the early stages of a design process, usability test tasks have a major role in refining the focus and scope of results.

2 Exploring the Unknown in Systems Development

Complexity, in systems development, can be regarded as characterizing either the context of system use or the designers' knowledge of it. The former is a natural part of the domain and hard to eliminate, yet often needs to be supported or partially resolved by technological artifacts. According to Mirel [1] such complexity is different in kind, not just degree, from well-structured simple tasks. Complexity from the latter perspective often characterizes the early phases of the system development. Traditional requirements engineering is based on the assumption that it is possible to recognise and plan for static requirements before actual development and

implementation work is started. Methods for conducting this type of engineering are well documented and validated and range from interviews and observations to prototyping in various forms [14]. In more advanced software development methodologies known e.g. as agile methods, the importance of formal planning is diminished and more direct communications between the actors are preferred [15]. It is admitted that the beginning of the design process is fuzzy [16], and the requirements are allowed to evolve in ongoing participation with future users.

Knowledge of users is hardly sufficient in terms of quantity and quality in many development projects. It is, however, impossible, even a design fallacy, to completely collect such knowledge [17]. Since the work by [18], it is known that the work practices during the system deployment may not correspond to the descriptions of work created and built into the system during the implementation process. The systems are exposed to drifting in usage and objectives [19], because users discover affordances [20], new and heterogeneous uses [21], and apply workarounds in order to get their work done [22]. Even ethnographic inquiries, which can address such intricacy of specific user contexts and work practices, are ineffective and insufficient in fulfilling the needs of design [17].

While the user involvement in design has become a truism for many IS development projects [23], the original, idealistic, picture of equal power of stakeholders in participatory design [24] has been reduced in user-centered methods to a power of system developer [25]. System developers and evaluators decide how much user participation is allowed [26] and what it means that a system is well-designed [25]. Thus, the user participation is institutionalized under the logic of technology development [25]. Likewise, the user practices are planned by the design before the actual use [27]. A true participation requires that potential users and stakeholders have a possibility to formulate and express questions and problems of the design and eventually have power to define the target of the development and possible solutions to it. An example of how to achieve this is to apply simple mock-ups in tests. These will evoke a variety of comments by users around the context, whereas with more detailed mock-ups the conversation focus is on the artifacts [28]. Discrepancies found in such simple solutions are inherently valuable in opening up the design decision-making and leading to an open-ended design process that extends into actual artifact use (cf. [27]). In contrast, a predefined scope of the development project and fixed interests of the developers will fundamentally ignore other possible perspectives and unforeseen parts of the solution.

2.1 The Role of Usability Testing and Test Tasks in the Design Exploration

Usability testing is a one of the most applied, evaluator-led, user inquiry methods under the umbrella of user-centered design. The method is well-established and seamlessly integrative with modern software development methodologies, yet regarded as a more confirming and disproving than exploratory and innovative method for designing ideas [29]. One of the reasons is that the evaluator rigidly controls what is asked, seen and provided – in good and bad [30]. This is manifested especially in the test tasks that are chosen by the evaluator and define in advance what type of results the

test will produce. Therefore the tasks and task formulations are critical in collecting usability data and both focusing the attention of evaluators and setting the scope for the whole test and its possible results. To succeed in the evaluation, evaluators need extensive domain knowledge [31]. Tasks and test scenarios are created based on knowledge of the domain and the product, its objectives, its target users and their supposed activities – knowledge, acquired in the previous development phases and requirements elicitation processes. A set of created tasks, as real and meaningful as possible, is further reduced based on various criteria (e.g. supposed frequency and criticality at actual use). In fact, domain knowledge is so important that domain experts without any usability expertise found more severe usability problems than usability experts in a study by [32]. Unsurprisingly, longitudinal user observations in the field revealed that testing the usability of system properties in the laboratory premises is not valid for real use situations [33]. The context-sensitiveness of usability work is frequently ignored in the complex health care domain too [34], where end-users' lack abilities to contribute to system development and the wide range of IT tools in the clinical context is overlooked [35].

Since the idea of UCD in the 80's [36] various method collections, method combinations and method modifications have emerged. Usability testing has also been harnessed to study users and practices at complex work domains. For example, [12] suggests conducting studies in the field of users, exploiting multiple evaluators, building simulations, developing situation awareness assessments, implementing unattended long-term data capture and using cued retrospective think aloud method with users. Another example is to apply a 'cooperative' usability test in order to gain knowledge about the work domain [37]. The essence of the method is retrospective interpretation phases, which ask why the user acted in a certain way and thus utilize user's knowledge of the work-domain to identify and understand usability problems [37]. In the field of health care, [38] combine an interaction sequence analysis to their usability study, pre-exploring the work practices of users before the tests with a contextual inquiry [39].

The weakness of classic usability testing is that test tasks may only concentrate on 1) features of the proposed design and 2) how these can be operated 3) in the known work tasks of the users the product is supposed to support. This leads to a premature commitment to the defined requirements [40], which rules out a wide spectrum of user requirements that might be needed [41]. Conventional usability test tasks, which include clear endings and correct answers, are not applicable to the analysis of complex systems [12]. Complex systems introduce usually much higher level goals than applied in typical usability testing tasks, which may be hard to specify beforehand while lower level usability testing may result in an easy-to-use solution for the wrong set of requirements [12].

Another explanation for this problem may be the existence of different and overlapping definitions of usability [42]. At one end, usability stops where utility begins [43], on the other end, usability is not a design issue at all, but inherent and inseparable from work and other goal-oriented action of people using tools (cf. [44]). In the latter perspective, one can be interested in how the system operates, yet subordinate the systemoperations to work processes that eventually determine usability and utility

of the system. That is not only about asking how well the system does (efficiency), but does it do the right thing at all (effectiveness). Usability studies with complex tasks have overlooked to measure the effectiveness aspect of usability [45]. Unfortunately, it may be easier to plan test tasks that are based on the current design than such that question it.

In order to increase their value the usability tests within design processes call for test tasks that fall outside the design solution and even outside the pre-specified context of use (Fig. 1). The expected benefit for the development process would be questioning and testing the design decisions made earlier (in Fig. 1 the unknown and known specified issues are presented as the nested circles of UCD process). That is bringing into discussion user needs and requirements, actual use situations and specific work practices that have not been included in the design solution and the intended sphere of usage. To define these tasks in practice, the potential user can, for example, bring her own tasks to the test session [41]. The benefit is that in addition to evaluating usability from the perspective of the design, the users' perspective is taken into consideration. Some of the user-defined tasks [41] may fall inside the proposed design solution, i.e. product functions support their execution while some may fall outside i.e. the tasks have not been implemented into the design.

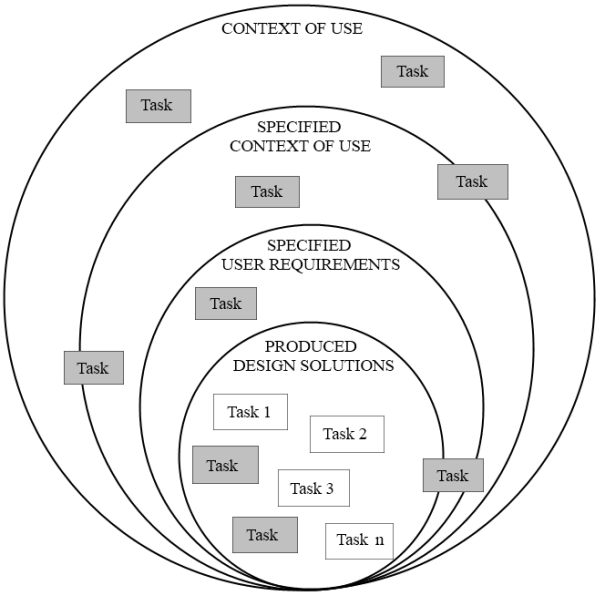

Fig. 1. Two different sets of test tasks: Pre-planned (1-N) and user-based (shaded tasks), which may fall anywhere between the context of use and the proposed solution (nested circles).

A modification of user-defined tasks, open-ended tasks, introduces another approach to empower users in a test session. In addition, such tasks cut the link with the assumptions made in the design process. An open-ended task approach does not point to any product function but to a whole work (process) inside the sphere of the context

of use. It is formulated as a high-level request for test participants to perform their work with the system under evaluation as a support. In this way, users can follow their natural flow of work and articulate their situational needs more freely [46]. Thus, instead of pre-defined models of work, users' situated work practices act as a starting point for usability evaluation and analysis.

Next, we discuss results that were acquired with open-ended tasks in three usability tests in the health care domain. The complexity of health care work is shortly explained and further discussed in the case-specific context.

3 Testing Usability in Three Complex Health Care Settings

Clinical work in the health care domain involves natural complexities [2], [47], [48]. Mirel [1], [49] relates these to complex problem solving, which is characterized by, among other things, vague goals, multiple methods and paths and lack of a distinct right answer. Handling these complexities of health care requires boundary crossing, polycontextuality and horizontal expertise [50]. For example, physicians face new problems in patient interventions that cannot be quickly turned into codified and repeatable procedures. According to Berg [51], clinical work is "*…characterized by the constant emergence of contingencies that require ad hoc and pragmatic responses. Although much work follows routinized paths, the complexity of health care organizations and the never fully predictable nature of patients' reactions to interventions, result in an ongoing stream of sudden events. These have to be dealt with on the spot, by whomever happens to be present, and with whatever resources happen to be at hand…"* Furthermore, the work is characterized by distributed decision making, by 'multiple viewpoints' and by its 'inconsistent and evolving knowledge bases' [51] The organization of health work mostly involves multiple stakeholders, the goals and preferences of whom may not be aligned [51].

3.1 Highlighting the Drifting Work Practices in Home Care

Home care involves many professionals from distinct disciplines who work in a cooperative and coordinated manner to provide care services for people living in their own homes. Home care workers describe the work as "*highly personalized caring labor that often seeps out of its formal boundaries into informal, unpaid activities*" [52]. In addition, the complexity of the domain is highlighted by the emergence of sudden and unpredictable care situations and needs due to less frequent monitoring of the patients.

In this domain, a mobile application was designed for nurses to be used during home care visits to clients. In order to perform patient visits, the nurses need access to client's contact information (where to go), care plans (what to do), and possibility to view earlier care treatments and actions (what has been done). This work process as well as an analysis of stakeholders, user profiles and use cases was represented by the developer in seven pages long document describing the context of use of the mobile application. In addition, user requirements of the new application design were based on knowledge about a desktop version of the application implemented earlier into the home care organization by the same developer. The new mobile application aimed,

however, to partly replace the desktop application. While the desktop application is currently used in the office premises, before and after the daily visits to inquire and entry clients' health data, the new tablet-based mobile application aimed to offer the same features and information during the visits at clients' premises.

A usability test for a paper prototype of the mobile application was conducted with four home care nurses working in two different units. The users were given an open-ended task as follows: "Your name is Z.Z. and you work as a nurse in the home care unit. Today is Tuesday 17th March. You are going to your third home visit of the day. The next client is called A.A. Please, perform your work and use this new application for support when needed." Due to data contained and presented in the paper prototype, the open task needed to be more detailed than originally planned i.e. the day, the name of the nurse and the name of the client have been fixed already by the paper slides and were unchangeable during the test. For the same reason, users needed to simulate their work in the test. Users were asked to think aloud while performing their client visit and documentary tasks with the artifact, and administrators intervened when necessary in order to understand the actions performed. Each session lasted 1.5 hours and was video recorded and transcribed.

Apart from some more minor issues, we found that every user had unexpected difficulties finding the purpose of the visit. The reason was lack of information about daily care tasks in the application. In nurses' terms this "daily information" is the most critical information to start the working day. The current work practice is that the daily information about each patient visited is entered as a free text to the desktop application at the end of the day. Next working day it is printed on the paper, carried along the day and fulfilled with new remarks about patients. The daily information serves various purposes. First, it is to inform the nurse what care actions are needed in the specific visit while preparing for visits. Second, it serves as a to-do list for future tasks (e.g. call the daughter on Friday/bring medicine next week) and a checklist for ongoing and past visits. Third, it is used to inform and communicate to other nurses about client related work. Fourth, and most importantly, the nurses conceive patients' current health situations through these entries of daily information.

It is notable that the daily information is a combination of three distinct text fields in the desktop application, yet managed as one set of information on the paper. All three fields were not implemented into or represented as a whole in the mobile version. The mobile application design was based on the assumption that the care plan functionality would be sufficient i.e. answer what kind of treatment the client requires. While the real information need was about the client needs for one particular visit, the care plans provided rather general and stable information. However, the original assumption of developers about the information content was not totally wrong, because the use of the care plan functionality on the desktop system had drifted during the years; the plans were not updated very often because they could not be accessed during the visits. The daily information on the paper had overridden the care plan, which had somewhat outdated data.

The open task applied in the usability test allowed us to distinguish and compare three distinct sets of work practices in the analysis of observations: a) the work practices with the current desktop application in use, b) the planned work practices im-

plemented in the mobile application and c) the actual work practices with the mobile application in the test session. The results were valuable for the further design of mobile application. Not only was more knowledge acquired about users and their activities with the mobile application, but about the foundation of these activities. Based on this knowledge, more justified design decisions could be made by both the developer and by the user organization: For example, whether it is necessary to keep the care plans updated and accessed during the visits i.e. to change the current work practices to benefit from the mobile application.

3.2 Identifying the Dual Roles of Users in Occupational Health Care

The aim of the development project was to redesign a current electronic patient record system (EPR) used by occupational health care providers. The first phase of the development was confined to functionalities on the physicians' and nurses' desktop module. The redesigned system was supposed to cover the whole care process and serve as a tool for nurses and physicians to carry out their daily work tasks. The initial requirements specification by the developer determined that with the desktop module physicians and nurses could manage appointments, health record entries, health measurements (blood pressure, weight etc.), laboratory examinations, prescriptions and customer invoices. The work of physicians is highly knowledge- and information intensive. It is hardly describable as general processes although routines based on legislation and evidence-based diagnosis practices exist. Complexities stem from patient interventions that require all the possible resources, information and tools at hand irrespective of the initial conditions and assumptions.

The first actual user research was carried out by a third party company renowned and specialized in industrial design. They interviewed seven end-users and observed their work practices and problems with the current software application in use. Findings of the user research were explicated and analyzed in 15 pages long report. This knowledge of users was supplemented and refined in cooperation with the domain experts of the developer organization. As a result, the design company created an initial wireframe of the EPR application.

A usability test for a paper prototype of the EPR system was conducted by the authors for two nurses and four physicians in the premises of the care provider. The users were given an open-ended task to "perform a patient visit and use the redesigned EPR when needed" i.e. to continue their work rather normally from the next patient as the sessions were arranged in the middle of the work day. However, the data in the paper prototype could not allow treating real patients during the test. Users were asked to think aloud while performing their work with the artifact and administrators intervened when necessary. Each session lasted about an hour and was video recorded and transcribed.

The most interesting finding in the tests was what the test subjects did after the patient visit. The nurses that participated in the test were operating the system very well, basically entering health data in a structured form to the system without major usability problems. However, nurses were far more interested in what kind of data they entered than how it was done. The reason was that they expected to fill in data that

can answer practical questions from the patients and the companies they serve. For example, a company that pays for the occupational health care may want to know "how many employees in our company have high blood pressure?" which questions the nurses are responsible to answer. Thus, we found out how the nurses not only had a role of a care worker but also an information broker and analyst, who compiles statistics for different well-being reports, communicates these to customers and frequently answers diverse health related questions.

In a similar manner during the test sessions, the physicians were highly focused on and expected to find the system functionality related to tracking the number of patient visits per day and the status of invoicing of each patient visit. The reason was that even the physicians had dual roles in the occupational health care. Firstly, they keep a common doctor's practice and care patients coming from the customer companies of the health care provider. The redesign of the EPR system was targeted at this role of the physicians. However, the physicians are also individual entrepreneurs who have their own business, which was not considered in the redesign. Physicians' business is performed in the premises of the care organization and with the tools and infrastructure, such as the redesigned EPR module, provided by the organization. Although the paper prototype introduced the customer invoicing functionality, the physicians worried whether their personal entrepreneurial requirements are implemented into the system and considered in the development. Due to the early stage of development and the scope of the project these requirements were not visible.

The usability test with the open task allowed discovering and defining different roles of the nurses and physicians of the system. While these roles were not discovered earlier or deliberately ignored by the development team, the user requirements of all relevant stakeholders were not present at the system specifications, which further indicated that the project scope of redesigning the EPR system was somewhat misaligned. Thus, a critical review of design project scope was a necessary action.

3.3 Exemplifying the Clinical Problems and Power Relations of National Health Archiving

The aim of the development project was to build a National Archive of Health Information, which is centralized data storage for health records that are accessed and used via local EPR systems. The basic idea was that the health data created in one local EPR system is stored into the national archive and can be later retrieved into the same or another EPR system in another health care organization. In addition, the operational logic of the archive required that health records and related management practices were standardized across the nation. The development of the archive was initiated in 2007 and the requirements specification was led by government institutions in an open and public way in cooperation with EPR developers and related stakeholders such as pharmacies and medical associations. Such a massive development project is very complex from the design point of view. Furthermore, the wide coverage of the project, i.e. the services of every health care unit and work of every health care professional whether in public or private organizations, naturally includes characteristics of a complex domain just as the health care does in general.

A usability test for a local EPR system integrated into the archive was conducted by the authors with six health care professionals (two physicians, two head nurses, a ward secretary and a home care nurse). These professionals had, among about 100 others, used the archive-integrated EPR system in real clinical work for three months' pilot period. The usability test was carried out two months after the pilot period had ended, in order to get explicit information about the experienced problems and their causes during the pilot. To our knowledge, the pilot and the succeeding usability test, carried out during 2012, were the first attempt to test the national archive use through the local EPR system. Despite the fairly long development process actual user testing was not technically possible earlier.

In the test session, the users were given an open task to "carry out their typical work tasks using the system", which was the same fully functional EPR system they had used earlier. Tests were arranged in the hospital premises and sessions lasted from one to two hours during which users were thinking aloud while simulating their frequent and common tasks with the EPR system. Test administrators intervened with additional questions yet no pre-defined test tasks were given to users. All tests were video recorded and transcribed.

The usability test identified extensive usability problems in concepts, vocabulary and terminology used in the national archiving. The users were familiar with professional conventions and the local agreements regarding the contents of health records, but were unable to adapt to the nationally defined standard vocabulary during the three-month pilot. The archive integrated EPR version demanded, for example, that the headings of health record entries as well as their order of appearance were nationally unified. All users experienced problems in creating record entries during the session and had experienced these during the pilot as well. The physicians had considerable difficulties in finding the latest health record entry (even their own fresh entry), which is a rather critical task and frequently performed before patient appointments in health organizations. The reason was that record entries were re-organized based on a new concept of a service event: The latest record entry could fall under an old service event and be buried in massive records. Thus, the centuries-old tradition of chronologically ordered health record was interrupted. In general, users had major difficulties in understanding how the concept of the service event should actually be applied - when a new service event should begin or end. The problem of opening and closing a service event arose because every record entry, health document, and even an act of accessing the health record, were to be handled under some specific service event. As it was a forced act by the system, the number of created service events was surprisingly high during the pilot period i.e. users bypassed the problem by creating a new service event instead of caring a patient and managing the record under an existing event. In addition to service events, the archive introduced many other new concepts (e.g. the phase of care, reason to access the record, the headings of record entries), which required radical changes to clinical work and were experienced problematic by the users in the session. For example, physicians almost lost their ability to create and read health record entries due to changes in terminology. In fact, during the pilot period the harmful effects of these new concepts were overridden by workarounds, default values and ignorance.

The usability test finally concretized the problems of national archive development experienced by the clinical practitioners. In addition to practical usability problems in the use of the EPR in clinical work, the test indicated that the clinicians lacked power in the national archive development to define concepts and rules that highly affected their daily work. The difficulties were experienced also by the local EPR developers, because they needed to follow and interpret the national system specifications, and moreover, lean on the user research done for that part. Thus, the causes of problems were far beyond the usability of a single EPR system connected to the national archive. The problems and recommendations of the test concerned workarounds for the user organization, system changes in the limits of national specifications for the developer organization, as well as requests for the national archive developer organization to empower all relevant stakeholders in the process and abandon the concept of service events. Although the test report was praised by the user organization, clinical practitioners as well as the EPR developers, the representatives of the archive developers, i.e. the most powerful stakeholders, refused to drop service events due to over six years of development of the concept that far. However, some misunderstandings between the EPR and archive developers in translating requirements into implementation details could be pinpointed and resolved, which led to system redesigning at the both local and national ends before new implementation and pilot iteration.

4 Discussion

In complex work domains, the "series of short, discrete scenarios" of classic, 'common industry format' kind usability testing are not appropriate [12]. Thus, the question arises how to test usability within complex professional systems development so that the results are useful in steering and informing the design process. Our approach with open-ended tasks provides a relatively low-cost solution to this problem. This was demonstrated in the complex field of clinical health care where the scope of usability work needs to be broadened [35].

Empirical tests with open-ended tasks can produce various results that benefit the design and development of health care systems. What characterizes these findings is that the issues found are essentially outside the sphere of the expected or specified use cases and contexts of use. In the first case, the findings indicated lack of knowledge of developers about users' drifting work practices and workarounds. In the second case, findings about users' adjacent and unrecognized work roles indicated a need to critically review and refine the whole project scope. In the third case, the new national standardization of patient information and clinical practices exemplified the low power of end-users and system developer in a nation-wide development project. In consequence, apart from suggestions about the designs themselves, even domain knowledge, design project scope and user organizations' practices could be brought into discussion.

These results are representations of different types of misfits that are frequently confronted in the use of organizational systems. For example, our case with the national data repository is a clear case of an imposed structure that causes issues on the

fundamental levels of system and organizational design [53]. The case in home care found functionality misfits [54], which lead to reduced utility and efficiency while role-based misfits, which imply mismatches between responsibility and authority [54], were present in the occupational health care development. As such misfits often have their roots in the deep structures of the system-organization interaction [54] and are hard to explore even with ethnographic methods, the open-ended testing approach appears very appealing in terms of efficiency and effectiveness. We assume that such extensive results could not have been found by testing tasks and scenarios built on the pre-specified assumptions of the user requirements and use contexts. This does, naturally, not mean that using the other methods of requirements elicitation and user involvement could be substituted by simple usability testing, but that the usability testing method can with benefit be used to validate and refocus the results of the other methods.

Practicing open-task testing is not only about posing one open question in the beginning and listening to users for the rest of the test, but requires an active role of the administrator. Of course, the fundamentals of the think-aloud technique apply to open-task testing also – people are different and they have effects on results and procedures. Perhaps the main strength in the procedure is that the administrator in open-task testing needs to learn what actions and operations users' work activities consist of. It is not only the relationship of humans and computers in interaction that become analyzed with open tasks, but the whole activity system [55], where equally relevant elements and targets of evaluation with actors and artifacts are the actions [44]. However, as Mirel [49] points out, complex work is not supported only by emphasizing actions but studying interactions between conditions, constraints and actions. Open-task test results are firmly tied to studying such activity systems, because users do, or simulate doing, their ordinary work actions involving real objectives and motivation. Therefore, questions in the test session are not limited to the open task only, but as it is difficult for users in many work contexts to articulate explicitly how they work, administrator's effort is needed before, during and after the tests to make the work visible. During the test, this may mean constant intervening by the administrator especially when work actions are simulated.

Compared with ethnographic [56] or contextual inquiry methods [39] for evaluation purposes and for designing systems for complex settings, we maintain that the open-task technique is a relatively low cost due to minimal preparing phase and short interventions although a systematic comparison of costs and resources have not been made. In addition, comparison between the results of predefined and open tasks has not yet been experimented and could not be conducted in the above described cases due to their industrial nature and practical purposes. However, many of the found problems may demand further investigation with the above methods and therefore the open-task approach is for them not a competitor, but a complement. Specifically from an evaluation perspective it is a technique to catch the most profound problems of the artifact early in the development while learning more about users and use contexts.

5 Conclusions

Usability test tasks are essential ingredients for the early system design process. Furthermore, tasks are fundamental to usability tests to make an impact on the design. In the development of complex systems, usability testing with the extended scope and open-ended structure as presented in this paper, delivers not only specific knowledge about the user interactions with the system, but can even reveal issues that, despite rigorous user research efforts, have been overlooked in the preceding phases of system development. The approach can disentangle the evaluation from previous design assumptions and share the ideal of participatory design where users are empowered partners of the design and evaluation. As demonstrated empirically, the approach will benefit the fuzzy and ongoing design process in the exploration of multiple and alternative contexts and future directions of early design for complex systems. The results of the presented case studies could be incorporated into the following design iterations in practice. Therefore, for usability practitioners, we suggest applying open-ended test tasks especially for testing systems in complex settings. Yet, user-initiated test tasks can be used with benefits even in other work domains and with different types of systems than discussed here (see [57]). Furthermore, we encourage technology developers and user organizations as well to acknowledge the wide spectrum of the possible outcomes of usability testing, some of which are not manageable by designers only but require attention and actions by managers at different levels and organizations. We want to maintain that the open-ended approach is not overriding the traditional type of testing with narrower scope and focus on the design solution. Instead, by increasing our understanding about the context, it also gives more credibility to such problem lists, severity ratings and design recommendations.

References

1. Mirel, B.: Usefulness: Focusing on Inquiry Patterns, Task Landscapes, and Core Activities, In: Mirel, B. (ed.) Interaction Design for Complex Problem Solving, pp. 31-63. Morgan Kaufmann, Burlington (2004)
2. Ferneley, E.H., Sobreperez, P.: Resist, Comply or Workaround? An Examination of Different Facets of User Engagement with Information Systems. European Journal of Information Systems, 15(4), 345-356 (2006)
3. Rittel, H.W., Webber, M.M.: Dilemmas in a General Theory of Planning. Policy sciences 4(2), 155-169 (1973)
4. Coyne, R.: Wicked problems revisited. Design studies 26(1), 5-17 (2005).
5. Gothelf, J.: Lean UX: Applying Lean Principles to Improve User Experience. O'Reilly Media, Inc (2013)
6. Hornbæk, K.: Dogmas in the Assessment of Usability Evaluation Methods. Behaviour & Information Technology 29(1), 97-111 (2010)
7. Wixon, D.: Evaluating Usability Methods: Why the Current Literature Fails the Practitioner. Interactions 10(4), 28-34 (2003)
8. Hertzum, M.: Problem Prioritization in Usability Evaluation: From Severity Assessments Toward Impact on Design. International Journal of Human-Computer Interaction 21(2), 125-146 (2006)

9. Molich, R., Dumas, J.S.: Comparative Usability Evaluation (CUE-4). Behaviour & Information Technology 27(3) 263-281 (2008)
10. Gray, W.D., Salzman, M.C.: Damaged Merchandise? A Review of Experiments that Compare Usability Evaluation Methods. Human–Computer Interaction 13(3), 203-261 (1998)
11. Hornbæk, K., Frøkjær, E.: Comparing Úsability Problems and Redesign Proposals as Input to Practical Systems Development. In: Proc. SIGCHI conference on Human factors in computing systems, pp. 391-400 ACM (2005)
12. Redish, J.: Expanding Usability Testing to Evaluate Complex Systems. Journal of Usability Studies 2(3), 102-111 (2007)
13. Greenberg, S., Buxton, B.: Usability Evaluation Considered Harmful (some of the time). In: Proc. SIGCHI Conference on Human Factors in Computing Systems, pp. 111-120. ACM (2008)
14. Paetsch, F., A. Eberlein, Maurer, F.: Requirements Engineering and Agile Software Development. In: 2012 IEEE 21st International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises, pp. 308-308. IEEE Computer Society (2003)
15. Fowler, M., Highsmith, J.: The Agile Manifesto. Software Development 9(8), 28-35 (2001)
16. Sanders, E.B.-N., Stappers, P.J.: Co-Creation and the New Landscapes of Design. Co-design 4(1), 5-18 (2008)
17. Stewart, J., Williams, R.: The Wrong Trousers? Beyond the Design Fallacy: Social Learning and the User. Handbook of Critical Information Systems Research, 195-237 (2005)
18. Suchman, L.A.: Plans and Situated Actions: The Problem of Human-Machine Communication, Cambridge University Press (1987)
19. Ciborra, C.: From Control to Drift: The Dynamics of Corporate Information Infastructures, Oxford University Press (2000)
20. Norman, D. A.: The Psychology of Everyday Things. Basic Books, New York, (1988)
21. Salovaara, A., Öörni, A., Sokura, B.: Heterogeneous Use for Multiple Purposes: A Point of Concern to IS Use Models' Validity. In: Pennarola, F., Becker, J., Baskerville, R., Chau, M. (eds.) Proc. of ICIS 2013 (2013)
22. Gasser, L.: The Integration of Computing and Routine Work. ACM Transactions on Information Systems (TOIS) 4(3), 205-225 (1986)
23. Bødker, S.: When Second wave HCI Meets Third Wave Challenges. In: Proceedings of the 4th Nordic conference on Human-computer interaction, pp. 1-8. ACM (2006)
24. Greenbaum, J. Kyng, M.: Design at Work: Cooperative Design of Computer Systems. CRC Press (1991)
25. Spinuzzi, C.: A Scandinavian Challenge, a US Response: Methodological Assumptions in Scandinavian and US Prototyping Approaches. In Proc. SIGDOC, pp. 208-215 ACM Press (2002)
26. Steen, M.: The Fragility of Human-Centred Design. PhD Thesis Delft University of Technology (2008)
27. Redström, J.: RE:Definitions of Use. Design Studies 29(4), pp.410-423 (2008)
28. Brandt, E.: How Tangible Mock-ups Support Design Collaboration. Know Techn Pol 20, 179 – 192 (2007)
29. Hanington, B.: Methods in the Making: A Perspective on the State of Human Research in Design. Design Issues 19(4), 9-18. MIT Press (2003)

30. 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)
31. Chilana, P.K., Wobbrock, J.O., Ko, A.J.: Understanding Usability Practices in Complex Domains. Proc. SIGCHI Conference on Human Factors in Computing Systems, pp. 2337-2346. ACM (2010)
32. Følstad, A.: Work-Domain Experts as Evaluators: Usability Inspection of Domain-Specific Work-Support Systems. International Journal of Human-Computer Interaction 22(3), 217-245 (2007)
33. Riemer, K., Vehring, N.: It's Not a Property! Exploring the Sociomateriality of Software Usability. In: Proc. International Conference on Information Systems (ICIS), pp. 1–19 (2010)
34. Yen, P.-Y., Bakken, S.: Review of Health Information Technology Usability Study Methodologies. Journal of the American Medical Informatics Association, 19(3), 413-422 (2012)
35. Kaipio, J.: Usability in Healthcare: Overcoming the Mismatch Between Information Systems and Clinical Work. Aalto University Publication series Doctoral Dissertations, 105/2011 (2011)
36. Norman, D.A., Draper, S.W.: User Centered System Design: New Perspectives on Human-Computer Interaction, L. Erlbaum Associates Inc. (1986)
37. Følstad, A., Hornbæk, K.: Work-Domain Knowledge in Usability Evaluation: Experiences with Cooperative Usability Testing. Journal of Systems and Software, 83(11), 2019-2030 (2010)
38. Viitanen, J., Nieminen, M.: Usability Evaluation of Digital Dictation Procedure - An Interaction Analysis Approach. In: USAB 2011. LNCS, vol. 7058, pp. 133-149. Springer (2011)
39. Beyer, H., Holtzblatt, K.: Contextual Design. Interactions, 6(1), 32-42 (1999)
40. Diaper, D.: Scenarios and Task Analysis. Interacting with computers, 14(4), 379-395 (2002)
41. Cordes, R.E.: Task-Selection Bias: A Case for User-Defined Tasks. International Journal of Human-Computer Interaction 13(4), 411-419 (2001)
42. 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)
43. Nielsen, J., Landauer. T.K.: A Mathematical Model of the Finding of Usability Problems. In: Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems, pp. 206-213. ACM (1993)
44. Nurminen, M.I., Reijonen, P., Vuorenheimo, J.: Tietojärjestelmän Organisatorinen Käyttöönotto: Kokemuksia ja Suuntaviivoja (Organizational Implementation of IS: Experiences and Guidelines). Turku Municipal Health Department Series A (2002)
45. Frøkjær, E., Hertzum, M., Hornbæk, K.: Measuring Usability: Are Effectiveness, Efficiency, and Satisfaction Really Correlated? In: Proceedings of the SIGCHI conference on Human Factors in Computing Systems, pp. 345-352. ACM (2000)
46. 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 7946, vol. 7946, pp. 91-106. Springer (2013)
47. Plsek, P.E., Greenhalgh, T.: The Challenge of Complexity in Health Care. British Medical Journal, 323(7313) 625-628 (2001)

48. Rouse, W.B.: Health Care as a Complex Adaptive System: Implications for Design and Management. The Bridge 38(1)17-25. National Academy of Engineering, Washington (2008)
49. Mirel, B.: Dynamic Usability: Designing Usefulness into Systems for Complex Tasks. Content and Complexity: Information Design in Technical Communication, pp. 233-261 (2003)
50. Engeström, Y., Engeström, R., Kärkkäinen, M.: Polycontextuality and Boundary Crossing in Expert Cognition: Learning and Problem Solving in Complex Work Activities. Learning and Instruction, 5(4), 319-336 (1995)
51. Berg, M.: Patient Care Information Systems and Health Care Work: A Sociotechnical Approach. International Journal of Medical Informatics, 55(2), 87-101 (1999)
52. Aronson, J., Neysmith, S. M.: "YOU'RE NOT JUST IN THERE TO DO THE WORK" Depersonalizing Policies and the Exploitation of Home Care Workers' Labor. Gender & Society, 10(1) 59-77 (1996)
53. Sia, S.K., Soh, C.: An Assessment of Package–Organisation Misalignment: Institutional and Ontological Structures. European Journal of Information Systems, 16(5), 568-583 (2007)
54. Strong, D.M., Volkoff, O.: Understanding Organization--Enterprise System Fit: A Path to Theorizing the Information Technology Artifact. MIS Quarterly, 34(4) 731-756 (2010)
55. Engeström, Y.: Expansive Learning at Work: Toward an Activity Theoretical Reconceptualization. Journal of Education and Work 14(1), 133-156 (2001)
56. Viller, S., Sommerville, I.: Ethnographically Informed Analysis for Software Engineers. International Journal of Human-Computer Studies, 53(1), 169-196 (2000)
57. Tarkkanen, K., Reijonen, P., Harkke, V. Koski, J.: Co-Constructed Tasks for Web Usability Testing. Proceedings of the IADIS International Conference on Interfaces and Human Computer Interaction (IHCI), pp. 79-86. IADIS Press (2013)