Wilson, S., Galliers, J. & Fone, J. (2007). Cognitive artifacts in support of medical shift handover: An in use, in situ evaluation. International Journal of Human-Computer Interaction, 22(1-2), pp. 59-80. doi: 10.1080/10447310709336955 <http://dx.doi.org/10.1080/10447310709336955>

Cognitive Artefacts In Support Of Medical Shift Handover: An In-Use, In-Situ Evaluation

Stephanie Wilson[1], Julia Galliers[1] and James Fone[2]

[1]Centre for HCI Design, City UniversityLondon
Northampton Square, London EC1V 0HB

[2]Framfab, 1 Naoroji Street, London, WC1X 0JD

Abstract

Technologies introduced to support complex and critical work practices merit rigorous and effective evaluation. However, evaluation strategies often fall short of evaluating real use by practitioners in the workplace and thereby miss an opportunity to gauge the true impact of the technology on the work. We report an in-use, in-situ evaluation of two cognitive artefacts that support the everyday work of handover in a healthcare setting. The evaluation drew inspiration from the theoretical viewpoint offered by distributed cognition, focusing on the information content, representational media and context of use of the artefacts. We discuss how this approach led to insights about the artefacts and their support of the work that could not have been obtained with more traditional evaluation techniques. Specifically, we argue that the ubiquitous evaluation approach of user testing with its reliance on think-alouds and observations of interaction is inadequate in this context and set an initial agenda for issues that should be addressed by in-use, in-situ evaluations.

1. Introduction

The HCI community has long recognised the value of investigating how users interact with artefacts. However, there is a continuing dichotomy between techniques for studying artefacts in everyday use in order to understand the work practice or as a precursor to design and techniques for evaluating artefacts as they are created during the design activity. In this paper we argue for a blurring of this distinction in order to develop effective techniques for in-use, in-situ evaluation appropriate to complex and life-critical work domains.

Work study techniques such as Contextual Inquiry (Beyer and Holtzblatt, 1998) and Task Analysis (e.g. (Diaper, 1989)) focus on understanding work as a first stage in the design of new artefacts or systems. While the approaches vary, they have in common an important emphasis

on investigating the work, including aspects such as the sequences of activities that people undertake to achieve their intent and the structure of artefacts used in the work. They are structured techniques with a clear focus and scope, defined by the methods they use and the modelling techniques they provide. The resulting descriptions of the work are generative: they are intended to inform and inspire the creation of new interactive artefacts. Workplace studies (Luff et al, 2000) yield a broader view of work. They use ethnographies and field studies to reveal rich and detailed accounts of how activities are accomplished in real work settings, providing fascinating insights into work in diverse domains such as control rooms on the London Underground (Heath and Luff, 1991) and the International Monetary Fund (Harper, 1998). These studies, with their openness to discovery, have also been used to influence the design of artefacts to support work, particularly collaborative work. However, evaluating the use of artefacts is not the primary concern of any of these work-focussed techniques. Instead this is the domain of usability evaluation techniques which take a narrower view, focussing on the artefacts used in work rather than the work as a whole. Usability evaluation is primarily concerned with evaluating new artefacts during an iterative design process and specifically considers the use of the artefacts (as opposed to their information content for example). The most popular evaluation techniques, whether user-based testing (e.g. Dumas and Redish, 1999) or expert-based inspection methods (Nielsen and Mack, 1994) such as heuristic evaluation, search for the usability problems that users will encounter when using an artefact to achieve specific goals so that they may be remedied in a subsequent design iteration. Quantitative 'measures' of the usability of the artefact may also be taken (Hornbæk, 2006). Usability evaluation techniques usually require the evaluator to make assumptions about the work of users and characterise that work in terms of a limited number of specific goals that individual users are striving to achieve; the search for usability problems and the measurement of outcomes are then limited to the narrow scope of the accomplishment of these goals and no attempt is made to consider the impact of the design on other aspects of the work.

We argue that this scope is too narrow in certain cases and that techniques for evaluating artefacts must sometimes take a broader account of the work following the approach of the work-focussed techniques. Artefacts that support complex and life-critical work pose a particular challenge for evaluation; the work they support is not readily replicated in lab-based evaluations, nor easily abstracted into a few specific goals. Hence, it appears unduly limiting for evaluation to collect only data about how individuals use the artefacts to accomplish a small set of predetermined goals and to examine that data solely for usability problems. We need to evaluate not just how artefacts support low-level, individual goals but their impact on the goals of the system as a whole. Not just how particular design features impact on use by an individual, but how the design as a whole impacts on the work of the system in positive or negative ways. In essence, we need evaluation "in the large" rather than "in the small". We

contend that it is not simply that we need to evaluate in-use, in-situ, but that we need new evaluation techniques for this purpose; the ubiquitous approach of user testing with observations of interaction and participant think-alouds will not suffice.

Our purpose in this paper is to present an argument for the importance of in-use, in-situ evaluation and to identify the kinds of use issues that can be revealed in such an evaluation as a step towards developing techniques that can be readily utilised by practitioners. We do this by means of a case study in which we summarise an evaluation of two cognitive artefacts in everyday use to support the work of a healthcare setting. These cognitive artefacts were physical objects that aided and supported human cognition (Norman, 1991) in the setting, although neither of them was an interactive computer system of the kind that is more commonly the focus of usability evaluation. Our evaluation was not a formal, structured usability evaluation driven by predetermined user goals. Instead, the data collection was ethnographic and the interpretation of the data was framed by the theoretical perspective of distributed cognition. These approaches are not new, for example see Hutchins' (1995b) detailed account of navigation on a ship, but are not generally regarded as the everyday tools of usability practitioners. Likewise, we are not advocating that practitioners should adopt the approach we followed in its current form; our purpose is to make a case for the importance of a broader evaluation for certain classes of systems and to reflect on our results to identify the kinds of issues that should be covered by such an evaluation. Hence, the goals of the work reported here are threefold: firstly, we demonstrate how a broader evaluation led to insights about the two cognitive artefacts in relation to the work that simply could not have been achieved with more conventional user testing; secondly, we use the results to critique current approaches to usability evaluation for in-use, in-situ evaluations; and thirdly we build on this critique to set an initial agenda for future development of in-use, in-situ evaluation techniques.

The remainder of the paper is structured as follows. We first review the current state of practice in usability evaluation, particularly user testing. In section 3 we introduce medical shift handover (the work practice that our two cognitive artefacts supported), describe the artefacts themselves and the study that we conducted. We then present key results from the evaluation to illustrate the rich details that in-use, in-situ evaluation has the potential to reveal, comparing this with the findings typically yielded by current evaluation techniques. We conclude with a discussion of the implications for the development of in-use, in-situ evaluation techniques.

2. Current Practice in Usability Evaluation

Usability work, including usability evaluation, is becoming increasingly integrated into the design and development practices of a range of organisations; we might say that it is becoming

institutionalised (Schaffer, 2004). This is a tribute to the campaigning efforts of usability advocates over many years, but it is also leading to a standardisation of practice in evaluation that may well limit its usefulness for certain classes of system.

Usability evaluation may either measure the usability of a system or scrutinise use of the system in order to identify real or potential usability problems so that they may be attended to in a redesign effort, or both. A range of expert- and user-based evaluation techniques have been developed, although many have remained tools for researchers rather than practical tools for practitioners. Current practice in usability evaluation is largely based on user testing where representative users undertake pre-specified tasks in laboratory, or pseudo-laboratory, settings and evaluators collect think-aloud and observational data of system use in pursuit of these tasks in order to identify usability problems and hence design flaws (e.g. (Krug, 2000)). The emphasis is on formative, diagnostic evaluation and test sessions are generally limited to one to two hours. Success on task, task completion times and user satisfaction ratings are among the more common measurements taken to supplement the diagnostic data. Given the continuing expansion of the Web, it is not surprising that much usability evaluation work is driven by the evaluation needs of websites and this user testing approach is a direct response to those needs. Laboratory-based user testing is a reasonable approximation to real world use for many websites: they are frequently accessed from home and office environments, the users' tasks can be achieved solely through use of the website and broader contextual factors are of little relevance, the website must be instantly usable the first time that the user encounters it and must therefore require minimal learning. It is fairly straightforward to simulate this sort of use in a user test by recruiting the right participants and setting them the right kinds of tasks. As the gap between the evaluation situation and the real world is relatively small for these cases, the ecological validity of laboratory-based user testing is likely to be high. Furthermore, user testing in the laboratory has the advantage of offering control over variability in test conditions, hopefully giving a reliable method.

In reality, practitioners are of course more innovative in tailoring usability evaluations to address their concerns and practical constraints than this sweeping generalisation would suggest and new techniques are emerging, for example, to support remote evaluation (McFadden et al, 2002) and to take advantage of technologies such as eyetracking (e.g. (Bojko, 2006)). Meanwhile, controversy continues to rage over issues such as how many users are required for user testing (Nielsen 2000), (Perfetti and Landesman, 2001), inconsistencies between evaluators in identifying and rating usability problems (Hertzum and Jacobsen, 2003) and the relative merits of testing with users versus experts (Fu et al, 2002). However, these debates on the details of approaches are in danger of missing the real limitations of usability evaluation as currently practiced.

There are systems whose real-world use context cannot be simulated so readily in a laboratory setting, leading to concerns about the ecological validity of evaluations. They include systems that have moved off the desktop into mobile, ubiquitous and wearable technologies and other, perhaps less glamorous, systems that are used everyday by people in support of their complex work practices. These are systems such as electronic patient record systems whose use facilitates work tasks, but where the work is not achieved solely through use of the system, that may require significant learning for effective use and whose use evolves with the work. In cases such as medical systems where the work is dynamic, complex and life-critical, the imperative for effective evaluation is great and so is the challenge of doing so in an artificial, 'out-of-use' laboratory evaluation.

While some argue for the need to conduct field-based evaluations, others question the benefits. Kjeldskov and Graham (2003) report a literature survey which revealed that 71% of evaluations of mobile devices were conducted in laboratory settings. Hertzum (1999) compares user testing in a laboratory against a workshop test where users worked in pairs on set tasks in a conference room setting and a field test in which users self-reported problems by telephone or on a test form. The test conditions were not directly comparable but the field test required careful management in order to yield useful data and appeared to identify fewer usability problems. Kjeldskov et al (2004) report a study comparing a 'realistic' laboratory evaluation and a field evaluation of a context-aware mobile device. Participants in the field condition were observed undertaking tasks similar to those set for participants in the laboratory condition. The same data (think-alouds, observation) were collected in both conditions and analysed for usability problems. The results suggest that there is little added benefit to evaluating in the field in this way; in fact the laboratory evaluation revealed more usability problems. We should perhaps not be surprised at such a result given that this approach to user testing was developed for the laboratory; the authors themselves state "Other methods for understanding use and interaction like ethnographic studies can most likely provide different perspectives on context-aware mobile systems use". Kaikkonen et al (2005) compare testing a mobile device in a "normal" laboratory setting against testing in the field and report little difference in the results yielded by the two tests: identical usability problems were identified in each condition, but the frequency was higher in the field condition. They conclude that the field is not necessarily the best place to evaluate a mobile device. Conversely, Kjeldskov and Skov (2003) demonstrate the value of realism in usability tests through a study in which they compare evaluations in laboratory settings with varying degrees of realism and with domain and non-domain users. They report differences in the problems identified in the settings and by the different users. The 'advanced' simulation lab, with its greater realism, facilitated the identification of usability problems not revealed in the 'normal' usability lab. Goodman et al (2004) claim it can be hard to use field studies "to obtain an objective evaluation of a device, determine its performance or gain hard

evidence comparing one device or method with another" and instead advocate *field experiments* for evaluating mobile devices. These are quantitative, experimental evaluations that are carried out in-situ as distinct from qualitative, ethnographic field studies. In essence, participants are set tasks and an experimenter follows them to note observations and take measurements including both conventional usability measures that concern use of the device (e.g task times and error rates) and other measures that are less to do with the interaction and more to do with the use situation (e.g. perceived workload and distance travelled).

Some have argued that the standard lab-based approaches to user testing do not translate well to the field, but their focus has tended to be on the practical difficulties of conducting the evaluation and collecting the requisite data in challenging field settings. For example Kaikkonen et al (2005) comment on the difficulties of using data collection techniques such as think aloud, video recording and observations in the field. It is our contention also that it is not sufficient to transfer lab-based evaluation techniques to the in-use, in-situ setting, but our concern is not so much with the practicalities of the techniques as with the limitations in the data they yield and subsequent issues revealed. By restricting the data to think-alouds, direct observations of interaction and users' self-reports, a valuable opportunity is missed. However, as Hertzum (1999) points out, while there is a proliferation of studies comparing different evaluation methods, little has been done to compare usability testing in the laboratory against real-world use in order to understand their relative strengths and limitations.

Although our interest is usability evaluation in general rather than evaluation specifically for healthcare systems, it is worth noting the current practice in this area also. Evaluation of healthcare technologies has been heavily influenced by the approach of 'randomised clinical trials' developed for medicines and medical devices where the emphasis is on measuring benefits in terms of clinical outcomes. Heathfield et al (1998) criticise both this emphasis on measurable "economic benefits and clinical outcomes" and the use of clinical trials as an evaluation tool, in part because of concerns regarding the external validity of the results (that is, are they relevant to real use situations) and in part because they offer little insight into how the technology may be improved. Similarly, Hartswood et al (2000, 2003) suggest that clinical trials have too narrow a scope for an adequate evaluation of healthcare technology because they fail to consider the broader contexts of doing work in real work settings. They argue that ethnography should have a role to explicate what they term the "lived work" of the setting. Hughes et al (1994) propose and exemplify ethnography having a role in evaluation "as a systematic means of monitoring systems in their use". The account in (Hartswood, 2003) of an ethnographic evaluation of a software tool to support radiologists in breast screening work is one example of the wealth of detail that can be revealed when using ethnography for evaluating in-situ and their discussion of how the results are distinguished from those of a clinical trial is

similar in spirit to the work reported here. We take this further by reflecting more generally on the nature of the issues that can be revealed in in-use in-situ evaluation.

3. A Study of Medical Shift Handover and Its Cognitive Artefacts

In order to investigate the opportunities afforded by evaluating in-use, in-situ once we have stepped back from conventional user testing, we report and discuss findings from an evaluation of two cognitive artefacts in a healthcare setting. The setting was a paediatric ward with 20 beds plus a high-dependency unit in a medium-sized, general hospital in the UK. The evaluation was undertaken as part of a broader study of the work of the ward that focussed particularly on medical shift handovers (Wilson et al, 2005); the two cognitive artefacts were summaries of the current 'state of the ward' constructed by the medical staff to support these handovers. It is outside the scope of this paper to discuss the work of medical shift handover in depth; rather, we use this particular case study to reflect on the value of in-use in-situ evaluation.

3.1 Medical Shift Handover

Care for patients in hospital is provided by a vast socio-technical system including people, information technologies, equipment, regulations and procedures. Provision of this care must be a continuous process: it must continue across boundaries of time as healthcare practitioners change shift and it must also continue across boundaries of space as patients progress from one clinical setting to another, for example from ambulance to Accident and Emergency (A&E) department, from A&E to admitting ward, etc. The transfer of responsibility at each of these points of discontinuity in time and space is clinical handover; the National Patient Safety Agency (NPSA) in collaboration with the British Medical Association (BMA) Junior Doctors' Committee have defined it as "The transfer of professional responsibility and accountability for some or all aspects of care for a patient, or group of patients, to another person or professional group on a temporary or permanent basis" (BMA Junior Doctors' Committee, 2004). Studies have shown that the risk of a breakdown in the work of any critical system is significantly increased at these kinds of transitions and that the consequences of breakdowns can be catastrophic (Lardner, 1996), (Patterson and Woods, 2001), (Shepard and Kostopoulou, 1999). Specific examples of wrong practice inherited and continued at transitions are cited by Grusenmeyer (1995), Wears et al (2003) and Patterson et al (2004).

Medical shift handover is the particular form of clinical handover that occurs between medical staff at shift change and effective medical shift handover makes a vital contribution to safe patient care (Wears et al, 2003). However, current practice varies from impromptu and

informal to regular and formal handovers. In spite of emerging guidelines (RCP, 2005) there is no standard approach to medical shift handover at present and the work of handover is typically supported by a range of ad hoc, locally evolved, artefacts such as handwritten notes, whiteboards and doctors' personal PDAs. The paediatric ward in our study had regular, formal handovers overseen by senior medical staff. There were three medical shift handovers each day and they took the form of dedicated handover meetings attended by all available members of the outgoing and incoming medical teams. The handover meetings were held in a side room off the main paediatric ward at fixed times each day. Prior to each meeting, one junior doctor from the outgoing medical team would be responsible for preparing a written summary of the information to be handed over; it was these summaries that were the focus of the evaluation reported here. In the meeting itself, the same doctor then verbally 'handed over' the information to the incoming team using the written summary as a reference. Every patient on the ward was handed over in order of bed number and this was followed up with more general ward information such as anticipated admissions. After the initial summary of each patient by the presenting doctor, there was generally some discussion of the patient and plans might be made for future tests, discharges etc. The written summary was passed to one of the incoming team after the handover meeting.

3.2 Cognitive Artefacts and Medical Shift Handover

According to Norman (1991), cognitive artefacts are physical objects that aid or enhance people's cognitive abilities. When we consider a distributed cognitive system (Hutchins, 1995a) (Hutchins, 1995b), cognitive artefacts are objects that form part of the system, supporting the cognitive processes that are distributed across individuals and mediating their collaboration. We observed a range of artefacts supporting the cognitive work of the paediatric ward, including whiteboards, patient notes, IT systems and a miscellany of paper artefacts. They were introduced or evolved to serve different purposes and this was reflected in their form and content. Two cognitive artefacts were central to handover and were the subject of the evaluation presented here: the "Handover Sheet" and the "Doctors' Book"[1]. Both were written summaries of key information to be handed over from the outgoing to the incoming shift in the handover meeting. They were also used as evolving representations of the work that needed to be done during the shift and, as such, were referred to and updated periodically throughout the shift. They provided a means by which information could be shared with others and care co-

[1] We use the term "handover summary" to refer to either of these artefacts.

ordinated both synchronously and asynchronously. They were central artefacts in the co-ordination of clinical care.

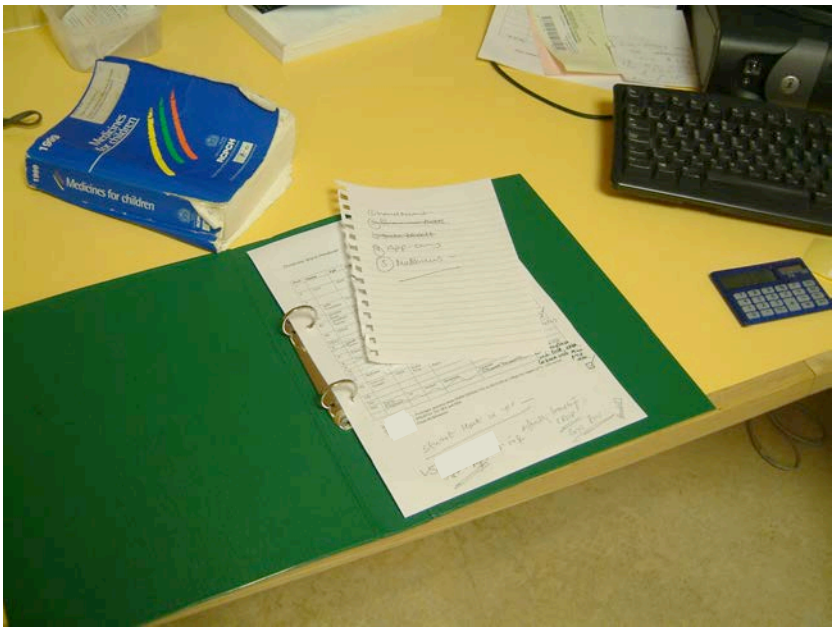

Figure 1: The handover sheet (anonymised)

The handover sheet was a print-out of an electronic document created using word processing software (Figure 1). The main part of the document was a table containing a row for every bed on the ward. For every occupied bed, the table included summary details of the patient (name, age, consultant, diagnosis, and jobs to be done). There was also some additional, general ward information, usually handwritten underneath the table (for example, forthcoming admissions or other anticipated events). The electronic document was continuously open on one of the computers at the ward nursing station and the medical staff would access and edit it periodically. Most notably this was done prior to handover when the presenting doctor updated the electronic document from handwritten notes on the paper print-out and other ward information resources. The doctor then printed a copy (the 'handover sheet'), put it in a green ring binder, and shredded the previous sheet. This was the only time during the shift when the electronic and paper versions of the document were guaranteed to be consistent. The ring binder containing the sheet was usually kept near the ward reception desk except during ward rounds when it would be on the notes trolley and during handovers when it was brought to the handover room. The handover sheet was readable, transportable and flexible, but notes made on the paper sheet during the shift were not necessarily transferred to the electronic document or to a more permanent record such as the medical notes. Updates during the shift were mostly written on the handover sheet, but were sometimes made directly to the electronic document (without a new version being printed). No historical record of previous handovers was retained because the electronic document was updated without back-up copies being made. The

handover sheet was in use when our study commenced but was later replaced by the doctors' book because risk managers in the hospital were concerned about breaches of confidentiality when copies of the sheet were left lying around the ward. They imposed the book as a solution to ensure that there was only ever one copy of the handover summary and that it would be easily locatable.

Figure 2: The Doctors' Book in the handover room

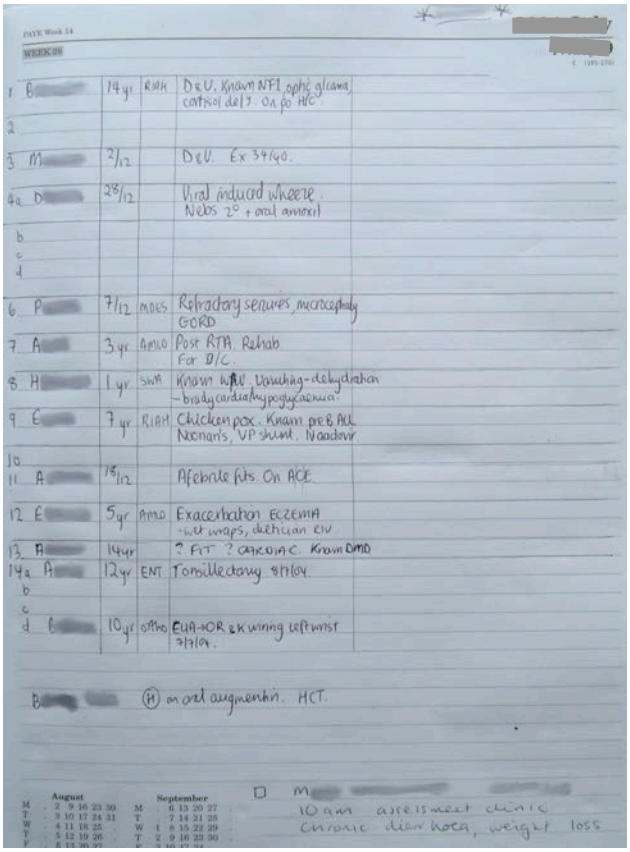

Figure 3: The Doctors' Book before first handover of the day (anonymised)

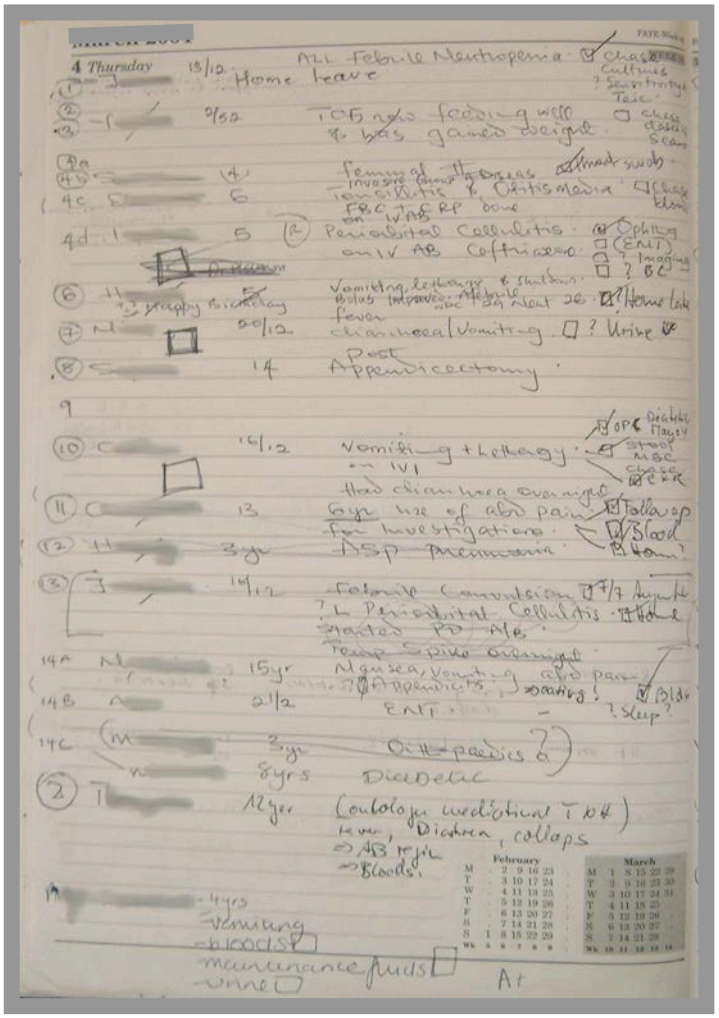

Figure 4 The Doctors' Book late in the day (anonymised)

The doctors' book (figures 2, 3 and 4) was a handwritten summary in a 'page-a-day' paper diary. It contained essentially the same information as the handover sheet: a list of patients in bed number order, their age, consultant, diagnosis and jobs to do. The first handover of the day required the presenting doctor to handwrite all the handover information on the page for that day, mainly by transcription from the previous day's entry. Later handovers on the same day involved updating the summary by handwritten edits to the original information. This representation was less readable than the handover sheet (because it was handwritten) and it became progressively messier during the course of the day due to the edits and the limited space in the diary. By the end of the day it was often very difficult to read. This artefact had the advantage of retaining a history of previous handover information (it was always possible to flick back to a previous page) and of allowing anticipated events to be entered for a future date, there was only ever one copy of it, it was easy to update and was flexible in terms of the information it could record.

Jones and Nemeth (2005) distinguish between what they term *exogenous* cognitive artefacts that are developed externally and then introduced to the workplace (e.g. a generic information system) and *endogenous* cognitive artefacts that are created by workers in support of their own work. Interestingly, these handover artefacts do not fall neatly into either category: they were both introduced by people other than those who used them on a daily basis (e.g. as mentioned above, the risk managers introduced the doctors' book to replace the handover sheet), but their detailed content and form was evolved by those who used them and, as a consequence, tended to vary slightly from day-to-day.

3.3 An Evaluation of the Handover Cognitive Artefacts

We conducted an ethnographic study of the medical shift handovers and their two supporting artefacts. The study was conducted over an elapsed time of 10 months during which we spent 24 days observing the work, including observations at weekends and in the evening. We investigated the work of the ward as a whole and medical shift handover specifically. We observed 37 different medical staff in handover (7 consultants, 13 registrars and 17 senior house officers). On average 7 medical staff attended the morning and evening handovers, at approximately 8.00am and 5.00pm respectively. There were fewer doctors at the 'late' handover which took place at approximately 8.30pm. The data collection was primarily observational, supplemented with opportunistic interviews, semi-structured interviews and detailed analyses of the two handover artefacts, similar to the data collected in the ethnographic study of operating room scheduling tasks reported in (Nemeth et al, 2004) (Jones and Nemeth, 2005). As far as possible, all aspects of the work that involved the cognitive artefacts were included: preparation for handover, handover meetings themselves, follow-up activities after handover, ward rounds and other updates to the artefacts during the course of the shift.

Direct observation: Detailed observations were conducted of the cognitive artefacts in the situations mentioned above. The researcher observed how medical staff used the artefacts in support of their individual and collaborative work, the information flows and resources in the setting as a whole, the physical and cultural context. Field notes were the primary form of data collection together with some audio recording in handover meetings. Video was considered too intrusive in this sensitive setting.

Opportunistic interviews: These spur-of-the-moment queries with medical, nursing and other healthcare staff supplemented the observations to elicit further information about the work and the use of the artefacts. They proved invaluable in enhancing the researcher's understanding of the domain: complex work is not readily understood from observations alone. Subjective views

of staff about the handover artefacts, the handover meetings and other aspects of the work were also obtained.

Artefact analyses: The two cognitive artefacts were observed to determine their information content (including accuracy), organisation, physical form, representations used (including symbols) and interactions. Field notes and photographs were used to record data.

Semi-structured interviews: Medical staff were asked detailed questions on a range of issues about the artefacts including their strengths and weaknesses, content, symbology, tasks and context of use. In addition, consultants (the most senior medical staff in the UK healthcare system) were asked as experts to rate the effectiveness of some of the handovers.

Clearly, our data was different to that obtained in conventional user testing and we treated it differently, adopting the perspective afforded by distributed cognition. Distributed cognition has previously been applied to understanding work in healthcare settings, for example, Xiao et al (2001) report using this approach in their study of a display board in an operating room display unit but provide few methodological details. We considered the distributed cognitive system of the paediatric ward and medical shift handover using notations developed as part of the DIB (Determining Information Breakdown) method (Galliers et al, In press) and evaluated the artefacts in this context. This involved firstly investigating and describing the information content of cognitive artefacts including the handover artefacts (e.g. bed number, name, age, consultant, diagnosis, jobs, anticipated admissions). We further examined all observed instances of use as recorded in field notes and photographs and the physical and organisational locations in which they occurred, the representational forms and interaction characteristics of the artefacts, information flows and breakdowns. We searched for system and individual goals (e.g. prepare for handover, maintain continuity of care, assess the state of the ward, preserve confidentiality) and the work activities in all their variations that contributed to the accomplishment of these goals. Finally, we examined the interviews with staff and the observational data to extract the positive and negative features of the artefacts and, crucially, compare and contrast their use.

4 Results

The evaluation of the two cognitive artefacts painted a rich picture of how they supported the work of shift handover and yielded many insights into their relative advantages and disadvantages. It is not the intent of this paper to provide an exhaustive report of the evaluation results but rather to use the results to illustrate the kinds of results that can be obtained from an in-use in-situ evaluation. In sections 4.1-4.5, we present examples of the (many) issues

identified in the evaluation for this purpose. We refer to them as *use issues* to distinguish them from *usability problems* and reflect their broader coverage.

First, however, it is important to note that the evaluation did reveal usability problems with both of the artefacts. Mack and Nielsen (1994) state that a usability problem is "any aspect of a user interface that may cause the resulting system to have reduced usability for the end user". In other words, a usability problem is any aspect of the design where a change would lead to an improvement on one or more usability measures such as effectiveness, efficiency, satisfaction, etc. Usability problems with the handover sheet included the difficulty of transferring handwritten symbols to the electronic document, the laboriousness of compiling the update from other information resources on the ward and the loss of information that occurred when completed jobs were deleted from the electronic document. For the doctors' book, the problems included the lack of space (especially on Saturday and Sunday) resulting in insufficient space for each patient, messy and illegible handwriting, the fact that the book was too cumbersome to carry around and, again, it took too long to write. Usability practitioners (e.g. Tognazzini (2001)) emphasise the importance of saying something positive about a system when reporting the results of a usability evaluation. However, perhaps not surprisingly, existing user-based and expert evaluation techniques uniformly focus on identifying the negative. We examined our data for the positive features of both artefacts. For the handover sheet these included that it was easy and fast to make notes on the paper copy, it was legible and it was transportable. The positive features of the book included that it also was fast and flexible to write on, symbols could be readily incorporated, it was locatable and completed tasks remained visible. However, because we did not have think-aloud data and had field notes rather than video of users interacting with the artefacts, some of the usual sources for identifying usability problems and positive features were not available to us. In all likelihood, we identified fewer of these issues as a consequence and there remains a case for addressing the issues more immediately associated with the user interacting with the system through the data collected in 'conventional' evaluation.

4.1 Location of update impacting quality of work

The first use issue we want to explore concerns the location where the handover summary (sheet or book) was created and its impact on the quality of the summary. The handover summary was compiled because no other artefact in the setting contained the information that had to be handed over in an appropriate form. Several artefacts each contained some of the relevant information. The handover sheet was updated on the computer at the nursing station on the ward. This was a central location where the most accurate information about the ward was available from information resources including the ward whiteboard, medical notes, patient

administration system (PAS) and other staff. In contrast, because the doctors' book was portable, it was updated not just at the nursing station but also in a variety of other locations including the canteen and the handover room. At first sight this portability seemed appealing: doctors welcomed the opportunity to have a coffee break while writing the summary. However, the portability actually appeared to have a negative impact on the quality and accuracy of the handover summary. This was because only a subset of the relevant information was available in these distributed locations; sometimes the only information resource was the doctor's personal knowledge of what had happened during the shift.

This highlights firstly the need to evaluate not just the process of interacting with an artefact and the users' subjective experience but also the quality of the outcome. This is in contrast to conventional user testing which tends to focus on difficulties during task performance rather than assessing the work that is achieved. In the case of complex work, it is not sufficient to determine task completion and quality by looking at use of the artefact in isolation. In this case, it was not enough to know that a doctor was able to complete the task of updating the handover summary and to measure how long it took to do so. We needed to reflect upon how well the task was done. For example, we looked at the accuracy and completeness of the information in the handover summary and asked experts to rate handovers. Clearly, assessing complex work is not a trivial issue, especially for non-domain experts, but this does not mean that we should ignore it. A simplistic approach in this setting would be to suggest that quality should be judged by patient outcomes. However this could only be measured over a much longer timeframe, it would be complicated by the many other factors in the complex reality of the work that also contribute to outcomes and it would ignore other important measures such as the effectiveness of the artefact in supporting other tasks (e.g. educating junior medical staff, providing a written record of the work), the efficiency with which the work was undertaken and the users' experience. Other measures of outcome that we are exploring include the effectiveness of handover as rated by the incoming shift and interrogating medical staff immediately after handover to determine how good a mental model they have acquired of the current state of the ward. There are parallels in other domains. For example, when testing an e-commerce website we should investigate not just whether the user is able to make a purchase, but how well that item satisfies their goals and constraints and Hornbæk (2006) in a review of current practice in measuring usability summarises a number of quality measures reported in the research literature.

Secondly, this use issue points to the importance of evaluating artefacts in the locations where the work is really done. Others have recognised the impact of context upon the *process* of using an artefact, for example (Goodman et al, 2004) comment on difficulties in using and evaluating

handheld navigation systems in adverse environmental conditions; in addition, we advocate explicit consideration of the impact of location on the quality of the outcome.

4.2 Updating as promoting a system check

Updating the handover summaries was laborious and there was the potential for error as the information had to be compiled from a variety of sources; this was particularly the case with the doctors' book where, for the first handover of the day, all the information had to be written by hand from scratch. However, looking at the bigger picture, our observations revealed that the update process also triggered what we termed a "checking the state of the ward" procedure: it triggered a checking mechanism for determining what was happening with each patient and whether or not required tests etc had been carried out. It was a fixed point in an otherwise busy shift where a review was required that might not otherwise have happened and where remedial actions were initiated. Further, as described above, the doctor needed to refer to and compare several information resources (including patient notes, ward whiteboard, nursing and medical staff) in order to create the summary. This acted also as a review of these resources. We observed instances where the update uncovered discrepancies between the information resources, (e.g. inconsistent spelling of patients' names, missing patients, out-of-date diagnoses, ambiguities regarding clinical tests planned or conducted) bringing to light confusion amongst the staff and initiating the resolution of these problems. Hence, creating the handover summary had a positive effect on the accuracy of other information resources in the setting, including the knowledge of the staff.

This use issue points to the need for an evaluation to attend to the real achievements of the work and the goals that transcend those of any individual. It was not sufficient to view the task of updating the handover summary as being only about changing the information in that artefact; it also instigated other cognitive processes in the setting, achieving a valuable update of other resources and encouraging a check on the state of the system, both of which have important implications for patient safety. This further reinforces the point made above regarding the importance of looking at the quality of the work achieved; in this case the quality concerns not just the handover artefacts but also other cognitive artefacts in the setting.

4.3 Retaining a history

Both handover cognitive artefacts were intended as memory aids for the doctor presenting handover to ensure that all relevant information was handed over to the incoming shift. However, they evolved to serve other, unanticipated purposes. One example of this was a consequence of the physical form of the doctors' book: it retained a (partial) history of the state

of the ward over the preceding days and months, enabling staff to refer back to earlier entries. This appeared to be done for a variety of reasons: to check factual information (patient details, tests etc), to discover why a patient had been admitted, to find out what had happened to a particular patient, to learn about what action had been taken in a similar situation etc. In general, it provided an archive that appeared to help medical staff construct a mental representation of the state of the ward. In contrast, the handover sheet did not retain any such history. It provided a snapshot of the current state of the ward. The electronic document was updated by editing the previous version without any back-up copy being taken and the previous paper sheet was shredded as the new one was printed. Any job on the sheet that had been completed was removed during the update (e.g. a patient discharge or test), leaving no record that it had occurred. Anyone else looking at the sheet could not then tell whether the job had been completed or whether it had been forgotten and omitted from the sheet.

The role of the doctors' book as an historical record is an example of how artefacts evolve to support unanticipated or unimagined tasks once they are introduced into work practice. Other examples we observed included staff using the handover artefact to communicate messages to each other and using it as a record of the current status of jobs to be done on the shift. This is not a new concept (see Carroll et al (1991) for example) but it is one that has remained outside the scope of mainstream usability testing. An in-use evaluation should not rely solely on assumptions about users' tasks that are made in advance of the evaluation but should recognise that tasks will have evolved and that these too should be included. In other words, an analysis of practice has an important role to play in in-use evaluation.

4.4 "It's only words!"

It was generally only the presenting doctor who had access to the handover cognitive artefact in the handover meeting: members of the incoming shift had to remember the information from the verbal presentation until the written summary was passed over afterwards. This was a significant problem especially for those who had not been on shift recently. As alluded to previously, medical shift handover is about helping the incoming shift to form an adequate mental representation of the current state of the system (the paediatric ward in this instance) so that they can assume effective control (see also Grusenmeyer (1995)). This internal representation need not be complete in itself but it needs to complement the external representation provided by the handover summary and this was difficult for staff to achieve when they did not have direct access to the summary. The problem was compounded by the fact that the practice of taking personal notes during handover was discouraged. This use issue was reported by staff who did not have access to the summary during handover, for example, one doctor reported that it was difficult to make sense of the handover information without a

written summary to refer to ("they're just words!"). This is an interesting example where the use issue is experienced by people who are not actually using the artefact. The problem is that certain aspects of the design of the artefact (in this case, single-user access) have a negative impact on the work of others in the socio-technical system. A further example of this was when staff who needed to read or update the handover summary were unable to locate the artefact because someone else had moved it to a non-standard location.

Once again, this is the kind of use issue that would be outside the scope of a conventional user test where the focus is on the person using the system. An in-use in-situ evaluation needs to consider the impact of the artefact on all agents in the system, irrespective of whether or not they interact with it directly. Hence it needs to go beyond collecting data about individual use. A further interesting point is that recent experience of the medical staff influenced the extent to which they encountered this use issue (i.e. time elapsed since last shift). It would have been difficult for us to pinpoint in advance the differences between staff that would impact their use of the artefacts but studying use made it clear that these differences did exist. This highlights the challenge for user testing of defining adequate participant recruitment criteria in advance.

4.5 Co-ordinating and communicating work

The handover cognitive artefacts acted as a resource for communication and co-ordination of the work between medical staff. They supported both synchronous and asynchronous collaboration activities. Most obviously, there was the 'formal' collaboration activity of passing the summary to the incoming shift at the end of the handover meeting. Staff were observed collaborating in updating the artefacts prior to the meeting and occasionally during it: sometimes someone would annotate the summary or request the doctor who currently had access to it to do so on their behalf. The artefacts also supported unintended forms of collaboration, for example, they became an important information resource throughout the shift for medical staff to determine at a glance the current state of the ward. Junior doctors wrote notes on the artefacts to record which jobs had been completed and which remained outstanding for their own benefit and to communicate the information to others, In addition to patient-specific information, the artefacts were used to communicate other ward information, providing a useful complement to the ward whiteboard. In some cases, poor handwriting, lack of space and inconsistent use of terminology and graphical symbols led to breakdowns in these communications. For example, if a blood test had to be carried out for a patient, this might be entered in the doctors' book with a checkbox drawn beside it. Ticking the checkbox was ambiguous: some people took it to mean that the bloods had been sent off for analysis; others assumed it meant that the results had been returned and dealt with.

While workplace studies have placed great importance on the kinds of collaborative activities summarised above, and hence have assumed an important role in the design of collaboration technologies, usability evaluation techniques have largely remained focussed on individual activity. An in-use, in-situ evaluation should investigate these collaborative activities as well as the individual work that is more commonly considered in evaluation, taking account of how artefacts support, and fail to support, both formal and informal collaborations.

5. Limitations of User Testing

The approach adopted in this study blurred the boundaries between the kinds of broad studies of current work practice that sometimes occur in the early stages of design or are undertaken by researchers in an endeavour to understand and describe the 'lived work' of systems and the focussed usability evaluations that are conducted later on. We sought both to evaluate the artefacts that supported the work and to understand the current work of the healthcare setting. The artefacts in our study were not sophisticated interactive systems, but the information they contained was complex and used in flexible ways to support and co-ordinate the work of the setting and to contribute to patient safety. Using ethnographic field studies to evaluate these artefacts in-use in-situ revealed their impact on the work of the setting as a whole and led us to use issues that could not have been obtained from either more traditional usability evaluations or from techniques such as task analysis which emphasise the cognition and activities of individuals and pay less heed to the use of artefacts in the work.

Conventional user testing of these cognitive artefacts would have involved asking representative users to undertake pre-determined tasks such as adding a patient to the summary, updating the jobs to be done or presenting a handover using the summary information; probably requesting the users to give a think-aloud protocol while performing the tasks and then observing and recording their use of the artefacts. It would have yielded data on individual use of the artefacts to achieve specific, pre-determined goals. The observational and think-aloud data would have then been examined to look for the problems the users encountered in interacting with the artefacts to accomplish the tasks. Clearly, we did not adopt this approach to evaluation; had we done so, we might have expected it to reveal detailed problems such as the poor legibility of the book and the laboriousness of creating the summary representation. In contrast, collecting ethnographic data and viewing it from the stance of distributed cognition encouraged us to focus on the goals and activities of the system as a whole (e.g. attend to patient safety through handover) and the contribution of the cognitive artefacts to the achievement of these goals. It revealed use issues such as how updating the artefacts promoted a 'checking the state of the ward' procedure and the importance of the location of the work.

This lends substance to the argument that it is not sufficient to conduct highly-structured user tests, whether in the laboratory or the field, especially of technology intended to support critical systems. It is only by conducting a broader study of artefacts in real use that the goals of the system as a whole can be discovered and these subtle aspects of the impact of the artefacts on the work can be revealed and understood. Based on the results of our evaluation, we summarise in Table 1 what a conventional approach to user testing focuses on (irrespective of where and when it is conducted) and what an in-use, in-situ evaluation could offer instead. Although we are making some broad generalisations, the overall point is not that people do not evaluate in-situ or in-use (they do sometimes, though not as often as they evaluate in-lab and before-use), but that conventional evaluations are not looking for the things that we would like to discover in an in-situ, in-use evaluation. While we have concerned ourselves primarily with user testing here, believe these limitations are true of most current approaches to formative evaluation.

'Conventional' User Testing	**In-Use, In-Situ Evaluation**
Examines the process of using an artefact	Examines the process of using an artefact and the quality of the work achieved
Evaluates from the perspective of individual goals	Evaluates from the perspective of individual and system goals
Relies on goals known at the outset	Includes new goals that evolve with the use of the artefact
Focuses on initial use of the artefact	Encompasses both initial and experienced use of the artefact
Examines individual use	Examines individual and collaborative use
Frequently conducted in artificial contexts	Conducted in the real use context
Searches for the negative (problems)	Searches for the negative and positive
Takes account of those using the artefact directly	Takes account also of those not using the artefact but impacted by its use

Table 1: A comparison of conventional user testing and in-use, in-situ evaluation

Criticisms of in-situ evaluations have pointed to the limited control available in the field as compared to the laboratory, claiming that this will lead to less reliable or robust evaluations. In response, we would point to evidence that laboratory-based evaluations are not as reliable as practitioners might hope. The CUE2 study (Molich et al, 2004) and the work of Hertzum and Jacobsen (2003) amongst others clearly demonstrate considerable variation in the results of evaluating the same system. Not only do evaluators vary in the details of how they approach testing, but different evaluators identify detect different usability problems and rate severity differently. A further counter argument is that we believe the value of the use issues revealed in in-use, in-situ evaluation is sufficient as to outweigh such concerns in the first place.

The use issues we highlight here were revealed in one in-use, in-situ evaluation. It would be valuable to conduct further evaluations of artefacts that support other forms of complex work in different settings, especially interactive artefacts, to determine whether the same kinds of issues arise and to search for others. It may also be fruitful to conduct a more formal comparison of conventional user testing against the kind of evaluation reported here. However this is methodologically challenging in a setting where the users have little time to offer, where it would be difficult to replicate the real work in artificial tasks for the lab and where the work is highly context-dependent.

6. Implications for In-Use, In-Situ Evaluation

In summary, there is clearly a role for in-use, in-situ evaluation as a complement to existing forms of usability evaluation. The things we would want to discover when evaluating an artefact in-use are in part different to the things we seek to discover prior to introducing it into use. We suggest that such an evaluation must firstly take a broader account of the goals/purposes for which an artefact is used. Specifically, it should consider how the artefact contributes to the accomplishment of high-level, system goals, both prescribed and non-prescribed, including those that evolve from the work practice. Secondly, it should investigate the ways in which people communicate and co-ordinate their activities to achieve these goals, and this means it must take account of those who do not use the artefact directly but are influenced by its use. Thirdly, the evaluation must examine not just the interaction process but also the quality of the work achieved and this should encompass both immediate and longer-term outcomes (e.g. both the quality of a handover summary and its impact on safe handover). Fourthly, the evaluation must take explicit account of the setting of the work and its influence on both the process of using the artefact and the outcome. Finally, it must focus on positive as well as negative impacts of the artefact on the work.

The use issues identified through attending to these concerns point to the consequences of specific design features. For example, the portability of the handover book, its ability to retain a history of handovers and the single-user access of both book and sheet were design features that had direct consequences for the work. In other words, just as usability problems identified in user testing lead us to reflect on design features and thereby attend to design flaws in redesign, so too can use issues identified in in-use in-situ evaluation lead us to design features and potential redesigns. Evaluating in-use in-situ has a role to play in the iterative design-redesign of artefacts: it is not just about measuring the impact of artefacts on work practice. We see this role as complementing rather than replacing conventional user testing: as mentioned above, think-aloud data is a valuable resource for identifying specific usability problems.

Usability practitioners rarely have the luxury of performing the kind of detailed ethnographic study that we undertook: the time and resources required are prohibitive. We are not advocating the method we adopted as a practical tool at this point in time. The challenge remains to develop new approaches to evaluation, approaches that blur the boundaries between work studies and evaluation, that will address the concerns raised here and reveal the kinds of use issues that we uncovered. A structured evaluation framework offers one possible way forward: our vision is that this would guide the evaluator both in collecting rich data about the use of an artefact and in looking in the data for the kinds of use issues articulated here. Our study suggests that the perspective afforded by distributed cognition offers one basis for this, in line with the vision of Hollan et al (2000) of distributed cognition as a new foundation for human-computer interaction. A distributed cognitive approach encouraged us to consider the cognition of the system as a whole, with its goals that went beyond the goals of any individual, information resources, flows etc. In (Galliers et al, 2004) we report an initial representational framework for describing these aspects that will provide the basis for further work. Further, it was fortuitous that we had the opportunity to study two different artefacts in support of the same work; two artefacts that at first glance might seem to have much in common, but whose differences had consequences for the work. Comparing and contrasting the similarities and differences in the artefacts and their relation to the work made an important contribution to the evaluation and allowed us to reflect on issues that might not otherwise have been apparent. In particular, it pointed to successes and failures in what otherwise appeared to be unremarkable use. This idea of comparative evaluation is one that we also wish to investigate further.

In summary, we have reported findings from an in-use, in-situ evaluation of two cognitive artefacts in order firstly to argue for the value of such an endeavour and secondly to reflect upon and articulate the issues (or some of them) that usability practitioners should seek to uncover during such evaluations and have discussed the shortcomings of current approaches to usability evaluation in this regard. In-use, in-situ evaluation offers the opportunity of revealing new

insights about artefact use and design flaws but different methods are needed to achieve this. This poses both a general challenge to develop new techniques to support the usability practitioner and a personal challenge for us in the future. We are investigating new technologies to support the work of handover but are faced with the dilemma that, on the one hand, we cannot introduce a new technology into real use in this critical environment without first evaluating its usability, while on the other hand we are aware that the true impact of such technology on the work can only be assessed in use.

Acknowledgements:

This work was funded in part by the ESRC-DTI-EPSRC (PACCIT) LINK Research Programme in the UK as the ACE (Information Appliances in Clinical Environments) project (ref: PACCIT RES-328-25-0002). We are grateful to our collaborators, the Bromley Hospitals NHS Trust, for the help they provided in this research.

References

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

BMA Junior Doctors Committee (2004) *Safe Handover: Safe Patients*, BMA.

Bojko, A. (2006) Using Eye Tracking to Compare Web Page Designs: A Case Study, Journal of Usability Studies, 1(3), pp. 112-120.

Carroll, J.M., Kellogg, W.A. and Rosson, M.B. (1991) The Task-Artifact Cycle. In *Designing Interaction: Psychology at the Human-Computer Interface*, Carroll, J. (ed.), Cambridge University Press, pp. 74-102.

Diaper, D. (ed.) (1989) *Task Analysis for Human-Computer Interaction,* Ellis-Horwood Ltd.

Dumas, J.S. and Redish, J.C. (1999) *A Practical Guide to Usability Testing*, Intellect Ltd.

Fu, L., Salvendy, G. and Turley, L. (2002) Effectiveness of user testing and heuristic evaluation as a function of performance classification, *Behaviour & Information Technology,* 21(2), pp. 137- 143.

Galliers, J., Wilson S. and Fone J. (In press) A Method for Determining Information Flow Breakdown in Clinical Systems, *International Journal of Medical Informatics*, Elsevier.

Goodman, J., Brewster, S.A. and Gray, P.D. (2004) Using Field Experiments to Evaluate Mobile Guides. In *Proceedings of HCI in Mobile Guides 2004 Workshop* (at Mobile HCI2004, Glasgow, Scotland).

Grusenmeyer, C. (1995) Shared Functional Representation in Cooperative Tasks – the Example of Shift Changeover, *International Journal of Human Factors in Manufacturing*, 5(2), pp.163-176.

Harper, R. (1998) *Inside the IMF: An Ethnography of Documents, Technology and Organisational Action,* Academic Press.

Hartswood, M., Proctor, R., Slack, R. and Rouncefield, M. (2000) Finding Order in the Machine. *Proceedings of the 21st European Annual Conference on Human Decision Making and Control*, Johnson, C. (ed), Glasgow, July 15th-16th.

Hartswood, M., Proctor, R., Rouncefield, M., Slack, R., Soutter, J. and Voss, A. (2003) 'Repairing' the Machine: A Case Study of the Evaluation of Computer-Aided Detection Tools in Breast Screening, Eighth European Conference on Computer Supported Cooperative Work, Helsinki, Finland, pp375-394.

Heath, C. and Luff, P. (1991) Collaborative Activity and Technological Design: Task Coordination in London Underground Control Rooms. In *Proceedings of the Second European Conference on Computer-Supported Cooperative Work*, pp. 65-80.

Heathfield, H., Pitty, D. and Hanka, R. (1998) Evaluating information technology in healthcare: barriers and challenges. BMJ, 316, pp. 1959-1961.

Hertzum, M. (1999) User Testing in Industry: A Case Study of Laboratory, Workshop and Field Tests. In Kobsa, A. and Stephanidis, C. (eds.), *Proceedings of the 5th ERCIM Workshop on User Interfaces for All,* Dagstul, Germany, pp.59-72.

Hertzum, M. and Jacobsen, N.E. (2003) The Evaluator Effect: A Chilling Fact About Usability Evaluation Methods, *International Journal of Human-Computer Interaction*, 15(1), pp. 183-204.

Hollan, J.D., Hutchins, E. and Kirsh, D. (2000) Distributed Cognition: Towards A New Foundation for Human-Computer Interaction Research, *ACM Transactions on Human-Computer Interaction*, 7(2).

Hornbæk, K. (2006) Current practice in measuring usability: Challenges to usability studies and research, *International Journal of Human-Computer Studies*, 64, pp. 79-102.

Hughes, J., King, V., Rodden, T. and Andersen, H. (1994) Moving Out from the Control room: Ethnography in System Design, In *Proceedings CSCW'94*, ACM Press, pp. 429-439.

Hutchins, E. (1995a) How a cockpit remembers its speeds, *Cognitive Science*, 19, pp. 265-288.

Hutchins, E. (1995b) *Cognition In The Wild*, MIT Press, Cambridge, MA.

Jones, P.H. and Nemeth, C.P. (2005) Cognitive Artefacts in Complex Work, In *Ambient Intelligence for Scientific Discovery,* Cai, Y. (ed.), LNAI 3345, Springer-Verlag Berlin Heidelberg, pp. 152-183.

Kaikkonen, A., Kekäläinen, A., Cankar, M., Kallio, T. and Kankainen, A. (2005) Usability Testing of Mobile Applications: a Comparison between Laboratory and Field Testing, *Journal of Usability Studies*, 1(1), pp. 4-17.

Kessner, M., Wood, J., Dillion, R.F. and West, R.L. (2001), On the reliability of usability testing, In *Proceedings CHI2001 Extended Abstracts*, ACM Press, pp. 97-98.

Kjeldskov and Graham (2003) A Review of MobileHCI Research Methods. *Proceedings 5th International Mobile HCI 2003 Conference,* Udine, Italy. Lecture Notes in Computer Science, Springer-Verlag, Berlin.

Kjeldskov J. and Graham C. (2003) A Review of MobileHCI Research Methods. In Proceedings of the 5th International Mobile HCI 2003 conference, Udine, Italy. Lecture Notes in Computer Science, Berlin, Springer-Verlag, pp. 317-335.

Kjeldskov, J. and Skov, M.B. (2003) Creating Realistic Laboratory Settings: Comparative Studies of Three Think-Aloud Usability Evaluation of a Mobile System. In *Proceedings 9th IFIP TC13 International Conference on Human-Computer Interaction, Interact 2003*.

Kjeldskov, J., Skov, M.B., Als, B.S. and Høegh, R.T. (2004) Is it Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field. In *Proceedings MobileHCI 2004 Conference*, Glasgow UK. Springer Verlag, pp. 61-73.

Krug, S. (2000) *Don't Make Me Think! A Common Sense Approach to Web Usability*. First Edition, Que Corporation.

Lardner, R. (1996) Effective Shift Handover – A Literature Review. Health and Safety Executive, Offshore Technology Report OTO 96 003.

Luff, P., Hindmarsh, J. and Heath, C. (2000) *Workplace Studies: Recovering Work Practice and Informing System Design*, Cambridge University Press.

McFadden, E., Hager, D.R., Elie, C.J. and Blackwell, J.M. (2002) Remote Usability Evaluation: Overview and Case Studies, *International Journal of Human-Computer Interaction,* 14(3&4), pp. 489-502.

Molich, R., Ede, M.R., Kaasgaard, K. and Karyukin, B. (2004) Comparative usability evaluation, *Behaviour & Information Technology*, 23(1), pp. 65-74.

Nemeth, C.P., Cook, R.I., O'Connor, M. and Klock, P.A. (2004) Using Cognitive Artifacts to Understand Distributed Cognition, *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans,* 34(6), pp. 726-735.

Nielsen, J. (2000) Why You Only Need to Test With 5 Users, Alertbox, March 2000, http://www.useit.com/alertbox/20000319.html.

Nielsen, J. and Mack, R.L (eds) (1994) *Usability Inspection Methods*, John Wiley & Sons.

Norman, D.A. (1991) Cognitive Artifacts. In *Designing Interaction: Psychology at the Human-Computer Interface,* Carroll, J. (ed.), Cambridge University Press.

Patterson, E.S. and Woods, D.D. (2001) Shift Changes, Updates, and the On-Call Architecture in Space Shuttle Mission Control, *Computer Supported Cooperative Work,* 10(3-4), pp. 317-346.

Patterson, E.S., Roth, E.M., Woods, D.D., Chow, R. and Orlando Gomes, J. (2004) Handoff strategies in settings with high consequences for failure: lessons for health care operations. *International Journal of Quality in Health Care*, 16(2), pp. 1-8.

Perfetti, C. and Landesman, L. (2001) Eight is Not Enough. http://www.uie.com/articles/eight_is_not_enough/

RCP (2005) Royal College of Physicians General Profession Training (GPT) Handbook www.rcplondon.ac.uk/pubs/handbook/gpt/gpt_handbook_app4.htm

Shepard, A. and Kostopoulou, O. (1999) Fragmentation in care and the potential for human error. *Proceedings First Workshop on Human Error & Clinical Systems (HECS)*, Glasgow Accident Analysis Group Report G99-1.

Shaffer, E. (2004) *Institutionalization of Usability: A Step-by-Step Guide*, Addison-Wesley Professional.

Tognazzini, B. (2001) How to Deliver a Report Without Getting Lynched, http://www.asktog.com/columns/047HowToWriteAReport.html

Wears, R.L., Perry, S.J., Shapiro, M., Beach, C., Croskerry, P. and Behara, R. (2003) Shift Changes Among Emergency Physicians: Best of Times, Worst of Times. *Proc. HFES 47th Annual Meeting,* pp. 1420-1423.

Wilson, S., Galliers, J. and Fone, J. (2005) Medical Handover: A Study and Implications for Information Technology. *Proc. Healthcare Systems, Ergonomics and Patient Safety (HEPS 2005),* Florence, Italy.

Xiao, Y., Lasome, C, Moss, J, Mackenzie, C.F. and Faraj, S. (2001) Cognitive Properties of a Whiteboard: A Case Study in a Trauma Centre. In W. Prinz, M. Jarke, Y. Rogers, K. Schmidt, and V. Wulf (eds.), *Proceedings of the Seventh European Conference on Computer-Supported Cooperative Work*, Kluwer Academic Publishers, pp.259-278.