Behaviour & Information Technology, Vol. 26, No. 5, September – October 2007, 421 – 436

Taylor & Francis
Taylor & Francis Group

Analyses of user rationality and system learnability: performing task variants in user tests

EFFIE LAI-CHONG LAW*†, BORKA JERMAN BLAZIC‡ and MATIC PIPAN‡

†ETH Zürich, Computer Engineering and Networks Laboratory (TIK),
Gloriastrasse 35, CH-8092 Zürich, Switzerland
‡University of Ljubljana, Institut Jozef Stefan, Laboratory for Open Systems and Networks,
Jamova 39, SI-1000 Ljubljana, Slovenia

No systematic empirical study on investigating the effects of performing task variants on user cognitive strategy and behaviour in usability tests and on learnability of the system being tested has been documented in the literature. The current use-inspired basic research work aims to identify the underlying cognitive mechanisms and the practical implications of this specific endeavour. The focus of our work was to assess user rationality and system learnability. The software application tested was a multilingual learning resource repository. Eleven German and eleven Slovenian participants were involved in two user tests (UTs). Usability problems (UPs) identified in two quasi-isomorphic tasks were categorized with respect to a scheme of associated skills. Actions of the two tasks of each of the 22 users were segmented and coded according to a scheme of cognitive activities. Results showed that generally the users adopted different strategies for working out the given task and its variant, and that the system could be proved learnable. User Rational Action Model and implications for future research on user tests are inferred.

Keywords: Rationality; Learnability; Usability evaluation; Task variant; Situated cognition; Mental models

1. Introduction

The current study is a use-inspired basic research (Stoke 1997) in a way that it is of theoretical importance to achieve the goal of understanding about user cognitive strategies underlying the initial and repeated usage of a software system, and also of the methodological importance to achieve the goal of use about the design of tasks for user tests and an alternative means to measure a system's learnability in terms of characteristics of usability problems. Indeed, there is a lack of empirical studies that systematically investigate these intriguing issues. Our work aims to bridge the gaps. Subsequently, we explore two key concepts of this topic, namely user rationality and system learnability.

The definition of usability put forward by Eason (1984) is rooted in the assumption that users are rational agents, interacting with a system by using their knowledge and deriving information from the system reactions to achieve their specific goals. According to Eason's causal model for usability, user knowledge, motivation and discretion interact with a cluster of task and system characteristics, leading to either a positive user reaction (i.e. the emerging strategy for the system use) or a negative one (i.e. the discontinuation of the use). In accord with user modelling techniques like GOMS (Card *et al.* 1983) and PUMA (Blandford *et al.* 2001), the usability of a computer system can be analyzed by observing how a cognitive architecture behaves when it is programmed with a particular scope of user knowledge. The key function of cognitive architecture is defined as providing and managing an agent's primitive resources that account for the agent's intellectual activities. In the traditional cognitive-rationalist paradigm, these

*Corresponding author. Email: law@tik.ee.ethz.ch

Behaviour & Information Technology
ISSN 0144-929X print/ISSN 1362-3001 online © 2007 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/01449290500484559

resources can be defined as the substrate upon which a physical symbol system (Newell and Simon 1976) is realized. In the situated-constructivist paradigm (Greeno *et al*. 1996), these resources can be defined as embodied psychological constructs (which are not necessarily symbolic) adaptable to particularities of a situation. In recognizing the complementary roles of the physical symbol system hypothesis and the situated cognition approach, we espouse the hybrid model that both situated and plan-based actions come into play when users work with an interactive system.

In user testing, which is one of the most widely applied methodological approaches for identifying usability problems of a system, test participants (i.e. representative end-users) are usually required to perform specific tasks with the system for which they have incomplete or even erroneous concepts. Negative transfer may occur as a result of applying incompatible mental models built from earlier interactions with similar artefacts to the current one at hand. We claim that these imperfect concepts can presumably be improved when users engage in achieving tasks with the system, given their ability to reason, learn and reflect. The improved mental models of the system will better support the subsequent interaction; users can then accomplish their tasks more effectively and efficiently. Besides, the nature and instance of usability problems identified during the initial and subsequent interactions with the system can vary. Note that usability problems can be attributed to users' defective mental models of the system as well as to the deficient design of the system.

According to the situated-constructivist view, people and cultural artefacts mutually shape each other's behaviour and development. How learnable should a system be so that novice users can adapt their mental models to situational demands with ease and effectiveness? Indeed, learnability[1] is one of the quality metrics to be evaluated in usability tests. Three well-known usability models, namely Eason (1984), Shackel (1986) and Nielsen (1993), commonly emphasize ease of learning but define it somewhat differently. Shackel's definition is comprehensive, including not only initial learning but also relearning. The former implies how easy it is for users to accomplish basic tasks the first time they encounter the design and how quickly a novice can become an advanced beginner, whereas the latter implies how easy it is to relearn an infrequently used application (cf. Nielsen's notion of 'memorability'). Note that in both cases some amount of training may be involved. Furthermore, in designing a system that users visit only occasionally (e.g. a brokerage system for learning resources; automatic teller machine), learnability becomes the most important element of usability. In contrast, when designing a system for power users (e.g. a corporate accounting system, CAD software) who need to use the system daily, learnability may sometimes be compromised for ensuring efficiency during use. Learnability can be measured in terms of subjective perceptions with the use of a retrospective questionnaire (e.g. SUMI; Kirakowski and Corbett 1988) and of objective performances in terms of task-completion-time and error rate. In the current study, we provide an alternative means of measuring learnability in terms of comparing the characteristics of the usability problems (cf. section 2.3) that have been identified in performing a given task and its variant.

In summary, in this paper we examine the effects of performing task variants on user behaviour with respect to their ability to reason and learn. To the best of our knowledge, there has not been any empirical study that systematically examines this specific topic. Results of the current study can provide better understanding of the cognitive mechanism underlying the observed effects and precise information about the tradeoffs in using task variants. Indeed, this study aims to advance the intellectual depth of usability evaluation—a young Research & Development field entailing deeper exploration of its theoretical foundation. Usability evaluation should be seen as an engineering process as well as a scientific endeavour (Gillan and Bias 2001).

2. Related works

2.1 *Effects of performing task variants*

In everyday educational practice using a single task to teach a concept is hardly ever the case: students are typically provided with one or more examples of how to perform a particular task, and are given a number of variants for drilling, application and consolidation purposes (Karasavvidis *et al*. 2000). Similarly, in a number of psychological experiments and research, the rationale of requiring participants to undertake task variants is to check the reliability of their performance or to reinforce their skills and knowledge to be acquired. Specifically, in the realm of usability evaluation, Lewis (1994) suggests that the likelihood of discovering a specific usability problem can be increased if users are required to perform the same task repeatedly or a task variant with a system. This proposition is somewhat consistent with the situated action view (Suchman 1987) that re-execution of an action frequently

[1]Here we differentiate two related but distinct concepts to avoid any confusion: (a) a system's learnability is defined as the ease with which one learns to operate a given system effectively. Normally it is measured in terms of the time or effort taken to get accustomed to the system and its operation and how easy it is to remember operational details; (b) a system's suitability for learning is defined as the extent to which the system can enable a user to learn specific concepts. This quality is primarily determined by the instructional design underlying the system. In this paper, we address only (a) but not (b).

uncovers problems of understanding, not just because the same terrain is re-considered, but because the terrain is seen differently during re-consideration. On the contrary, performing task variants can improve user mental models about a system's features; they may also circumvent the usability problems experienced earlier by some work-arounds. Consequently, because of the practice effect both the number of usability problems and task-completion-time for working out quasi-isomorphic tasks may be reduced. Observing and analyzing differential user behaviours when performing a task and its variant can provide empirical evidence to resolve the above controversy—*whether more or less usability problems can be identified in performing a task as compared with performing its variant*. Besides, this empirical approach can serve as a practical means to evaluate the learnability of a system and to estimate user rationality and adaptability.

2.2 *Patterns of cognitive activities*

The rationality of a cognitive architecture is a measure of consistency and reasoning power. Generally, if an agent would perform two different actions with the same knowledge and goal in two identical situations, it is *not* said to be rational (Lemon *et al.* 1994). Conversely, if an agent's knowledge improves as a result of learning, it is expected to perform differently to achieve the same goal in the same situation. Allen Newell's (1990) principle of rationality states: 'if an agent has knowledge that one of its actions will lead to one of its goals, then the agent will select that action'. This formulation implies that there is a direct connection between goals, knowledge and subsequent actions, and that rationality presupposes self-consciousness and reasoning ability of an agent.

In user tests, the goals of individual tasks are normally explicitly specified, although users may interpret them differently. We hypothesize that when entering an interaction with a system, novice users may attempt to employ means-ends planning—a weak problem-solving method—to perform the given tasks. However, given their incomplete mental model of the system, they cannot anticipate courses of actions or their consequences, and their rational approach is deadlocked. Hence, in the initial interaction with a system, novice users tend to engage in exploratory actions and opportunistic trials (cf. Rosson and Carroll 1995), which incidentally augment the solution space and enhance their conceptual knowledge of the system. Opportunistic trials can be seen as approaching the initial interaction barrier by means of selecting variants of the rational approach. Further, the exploratory actions are predominantly *ad hoc* or, in Suchman's (1987) terms, situated in the sense that users rely on their embodied cognition, which is built upon their past learning experiences in interacting with technical artefacts, to deal with different kinds of impasses or predicaments. This conjecture is consistent with Wilson's (2002) notion of 'representational bottleneck'. Accordingly, when users work under situations that demand fast and continuously evolving responses, they may simply have no time to construct a full-blown mental model of the environment, from which a plan of action can be derived. Instead, they generate situation-appropriate actions on the fly. Furthermore, we claim that owing to inadequate mental models, novice users are more likely to be perturbed by non-anticipatory reactions of the system. However, in the subsequent interaction with the system when these users become experienced but are not yet able to automate the skills required, their actions tend to be plan-based in the sense that they can best orient themselves to the resources in the environment (e.g. cues and feedback of the system).

2.3 *Categories of usability problems*

If people behave rationally, given that their knowledge and goals remain unchanged, they will persistently commit the same errors. It is claimed that the occurrence of systematic user errors is a side-effect of the user behaving rationally (Curzon and Blandford 2000). However, if users' mental models can be adapted as a result of their previous exposure to similar situations, then the number and nature of problems they experience are supposed to vary accordingly. In fact, human errors in interactive systems can be just as disastrous as device errors. Fu and his colleagues (2002), grounded in Rasmussen's (1986) theory of mental model, developed a classification scheme that associates usability problems with three levels of user cognition, namely skill-/rule-/knowledge-based. Accordingly, usability problems can be categorized based on the immediate cause of errors and the related information processing.

None the less, there is a major difficulty with Rasmussen's schema: the words 'skill', 'rule' and 'knowledge' are semantically rich and ascribed with different meanings by different people in different contexts. Specifically, Rasmussen uses the word 'skill' in a limited sense, referring to perceptual motor skills that are automated and require no conscious monitoring. He refers rule-based behaviour to so-called familiar cognitive skills (Bainbridge 1997) which arises in repeated task situations, when a person has become familiarized with a task through practical experiences and thus developed a standard method to accomplish it. Furthermore, Rasmussen's knowledge-based behavior is consistent with problem-solving skills, which are deployed to re-structure mental models to meet situational requirements. To avoid the inherent shortcoming of the words skill/rule/knowledge, we adapted the model of Fu *et al.* by specifying the three levels of cognition with the respective types of skills (table 1).

Table 1. Categories of usability problems and cognitive user models.*

Level of user cognition	Type of information	Types of errors	Categories of usability problems	Implications for design
Perceptual-motor skills	Signals	Schema activation; signal misperception; motor variability	Sensory modalities (perceptual, motor); feedback; attention	Provide percepts with strong affordance
Familiar cognitive skills	Signs	Rule omission; step omission	Cueing; consistency	Provide salient cues for error prevention
Problem-solving skills	Symbols	Formulation of incorrect intention; disorientation in problem space	User help; learnability; mental models; functionality	Enable development of mental model

*Adapted from Fu *et al.* (2002).

2.4 *Transfer of task knowledge*

Learning occurs when users acquire knowledge while working on a given task. Transfer of learning[2] occurs when users are able to apply the knowledge acquired to deal with a similar or novel task. Transfer includes *near transfer* to closely related contexts and performances and *far transfer* to rather different ones (Perkins and Salomon 1994). In our study, given the high similarity between the task and its variant and the short time lapse that allows the working memory to remain effective, near transfer of knowledge is likely to take place. The concomitant question is what kind of knowledge is involved. In cognitive psychology knowledge is categorized into two kinds with distinct characteristics (e.g. Johnson *et al.* 1998), which can be described specifically in HCI terms (Polson and Kieras 1985). Declarative knowledge is a user's knowledge of tasks that a system performs, the contexts in which tasks are performed, and how tasks are interrelated. Procedural knowledge is the knowledge of the actual operating procedures for a system and of methods used to perform tasks. The two types of knowledge are represented in form of schemata and production rules, respectively, in a user's mental model (Anderson 1993). In the current study, presumably due to direct hands-on experience with the system, the users' procedural knowledge is strengthened to a greater extent than their declarative knowledge. Hence, the transfer effect of the former is supposed to be stronger. Note, however, that the current study does not directly address the issue of transfer since we have not systematically manipulated the variables of interest (e.g. the task order, the degree of similarity between the two tasks). Nevertheless, some of the observations can be interpreted under the transfer paradigm.

[2]Transfer, being an age-old theoretical and practical problem, has fallen in and out of the central focus in the history of psychological research. In the mid-1990s heated debates on transfer were instigated by the Situated Cognition movement (Gruber *et al.* 1999). However, detailed descriptions are beyond the scope of this paper.

3. Research hypotheses

Subsequently, we use the terms 'task variant' and 'quasi-isomorphic task' synonymously to refer to two tasks that are very much like each other, but only a finite number of elements prevent them from being isomorphic or identical. Formally speaking, the degree to which two tasks are isomorphic is determined by the number of common elements and the number of common relationships between the elements. Isomorphic elements may have the same or different names (cf. phenotype), but they must represent the same object having the same properties and definition (cf. genotype). Basic elements of a task include {objective, input, operation, and output}. Each of the four elements can be divided into sub-elements. The threshold (i.e. percentage of the elements that are common) at which two tasks are classified as quasi-isomorphic or different is somewhat arbitrary, depending on the ultimate purpose of demarcating the two types. None the less, we propose that the threshold value is reasonably set to the minimum of 80% (cf. Cronbach's alpha measure of above 0.8 is interpreted as good correlation). In other words, the two tasks should share at least 80% of their elements to be named as quasi-isomorphic. On top of it, another critical criterion is that the two tasks must share the same generic goal (e.g. to create a record in a database system), although the specific objective may slightly be different (e.g. the record is in the form of educational material or educational activity).

Based on the foregoing arguments, we formulate the following hypotheses (**H**).

H1: When two quasi-isomorphic tasks sharing the same generic goal are performed serially, the mean Task Completion Time (TCT) of the earlier task will be significantly *longer* than that of the later one, because of the user's better knowledge state and thus more effective actions for the later task.

H2: When two quasi-isomorphic tasks sharing the same generic goal are performed serially, the number of Usability Problems (UPs) experienced by a user in the

earlier task will be significantly *more* than that of the later one, because the user is able to develop work-arounds to circumvent some UPs.

H3: When two quasi-isomorphic tasks sharing the same generic goal are performed serially, the number of errors and instances of help seeking observed in the earlier task will be significantly *more* than those observed in the later one, because of the user's better understanding of the system.

H4: The perceived ease and the perceived efficiency in performing a task are significantly *less* than those in performing its variant subsequently; these subjective measures are significantly correlated with the objective measures.

H5: The pattern of rational cognitive activities that a user exhibits when performing a task is different from that when performing its variant subsequently: the former comprises more exploratory and situated actions as well as more perturbations and repairs, whereas the latter comprises more planned-based actions.

H6: There will be a significant negative correlation between the user's self-reported expertise in the domain-specific knowledge relevant to the system tested and the user's tendency to engage in exploratory actions.

H7: Types of usability problems that a user identifies when performing a task are different from those when performing its variant subsequently, with more of the problems being associated with problem-solving skills in the former and more with perceptual motor skills and familiar cognitive skills in the latter.

4. Usability tests

4.1 *Design*

The system evaluated was a multi-lingual learning resource repository (LOR) enabling the exchange of online educational content among higher education institutions (http://www.educanext.org/ubp). Different versions of the user interface of the LOR have been usability tested for research as well as practical purposes. User tests on the earlier version (v. 0.9)—**UT1**—and User Tests on the current version (v.1.0)—**UT2**—were conducted in two academic institutions, one in Switzerland and another in Slovenia. Standard user test procedures were adopted (Dumas and Redish 1999). UT1 and UT2 were administered by the respective local experimenters, who were also responsible for recording the data and transcribing thinking aloud protocols of the participants. The user interface was originally developed in English and translated into different European languages, including German and Slovene. None the less, the English instead of the respective native language version was tested for the following reasons:

- to minimize usability problems that are caused by any translation error;
- to ensure the highest possible level of uniformity of the testing procedure across the two sites;
- to ensure the consistency of data analyses being performed by a usability specialist who cannot speak both languages; and
- to contain the resource in the budget as translating the test instructions and task scenarios, which were designed by the usability specialist and presented in English, can be very time- and effort-demanding.

Nevertheless, to compensate the drawback of using a non-native version, one of the criteria for recruiting test participants was that they should be fluent in English, both spoken and reading comprehension.

4.2 *Participants*

For UT1, seven (E1, . . . ,E7) and three (S1, . . . ,S3) participants from Switzerland and Slovenia were involved. For UT2, four (E8, . . . ,E11) and eight (S4, . . . ,S11) participants from the same two universities were involved. Altogether there were 22 of them. They were researchers, project managers, system developers, administrators, university professors and librarians. Their heterogeneous levels of competence in information technology and e-learning can partly account for the diverse usage behaviors observed.

4.3 *Tasks*

In UT1 and UT2, each participant was asked to perform ten task scenarios; all except one of these tasks were the same for both UTs. The nine common tasks covered the core functionalities of the LOR (table 2).

Additionally, in UT1 the participants were required to delete the learning resource provided whereas in UT2 the participants were required to design a task with a specific

Table 2. Nine common tasks of the usability tests (UT1 and UT2).

Task 1	Apply for a User Account
Task 2*	Provide and Offer a New Educational Material
Task 3	Modify the Discipline of the Educational Material
Task 4*	Provide and Offer a New Educational Activity
Task 5	Modify the Schedule of the Educational Activity
Task 6	Modify the Offer of the Learning Resource Provided
Task 7	Browse the Portal Catalogue
Task 8	Search Learning Resources in the Portal Catalogue
Task 9	Book and Access Selected Learning Resources

*Note: Task 2 and Task 4 are quasi-isomorphic tasks.

goal on their own. Specifically, Task 2 (Provide and Offer a New Educational Material) and Task 4 (Provide and Offer a New Educational Activity) are highly similar, sharing a number of common metadata attributes that should be filled in by users in four basic steps, namely 'General (1)', 'General (2)', 'Technical' and 'Educational' (see the cascading windows in figure 2). Task 4 consists of an extra step 'Scheduling' for specifying a timetable of a live educational activity. Note that the content (i.e. metadata attributes and workflow) of the core functionality 'Provision and Offer of Learning Resources' was basically the same for v.0.9 and v.1.0, but there were some obvious changes in the presentation (cf. figures 1 and 2).

4.4 *Procedure and instruments*

In both UTs participants were escorted into a testing room and seated at a desk with a computer system (Model: PC; Operating system: Windows XP; Browser: MS Explorer 6.0; Networking: LAN or T1). They were asked to complete a pre-test on background data, an 'After-Scenario Questionnaire' (Lewis 1995) for each of the ten tasks, and a post-test 'Computer System Usability Questionnaire' (Lewis 1995). The participants were asked to maintain a running commentary as they interacted with the system. Such verbalizations together with screen activities were captured by a specific software application (Camtasia® or Hypercam®). Individual task was recorded separately with a unique filename (e.g. S1_task1). The recording was initiated manually either by the local experimenter or by the participant at the moment when she finished reading the instruction and was ready to carry out the task with the computer; the recording was manually stopped when the participant explicitly stated that she completed or quitted the task. The task completion time (TCT) was computed by the difference between the starting-time and finishing-time. The average TCT over the nine common tasks was 45.9 minutes and 33.8 minutes for UT1 and UT2, respectively.

5. Data analysis

A usability specialist transcribed, segmented and analyzed the video-recordings of Task 2 and Task 4 for each of the 22 participants. Two criteria for segmentation are: (i) a change of user sub-goal underlying the observed action, which was explicitly verbalized in thinking-aloud or inferred from the participant's performance by the usability specialist, and (ii) a change of the system status as a result of the observed action. Besides, when analyzing an action the immediate preceding and succeeding actions are taken into consideration to attest the associated sub-goal. The segments so derived were coined as 'action segments'. Each segment is coded based on the following scheme (table 3). An example of action segment analysis is presented in Appendix A.

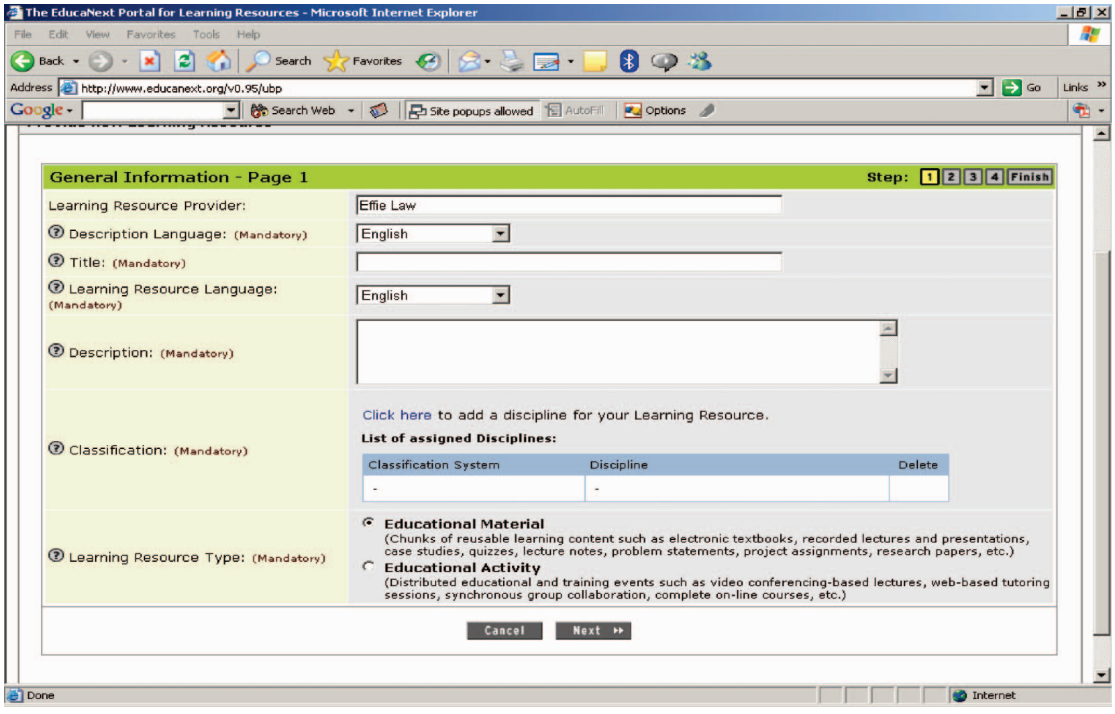

Figure 1. The academic portal usability tested—v. 0.9.

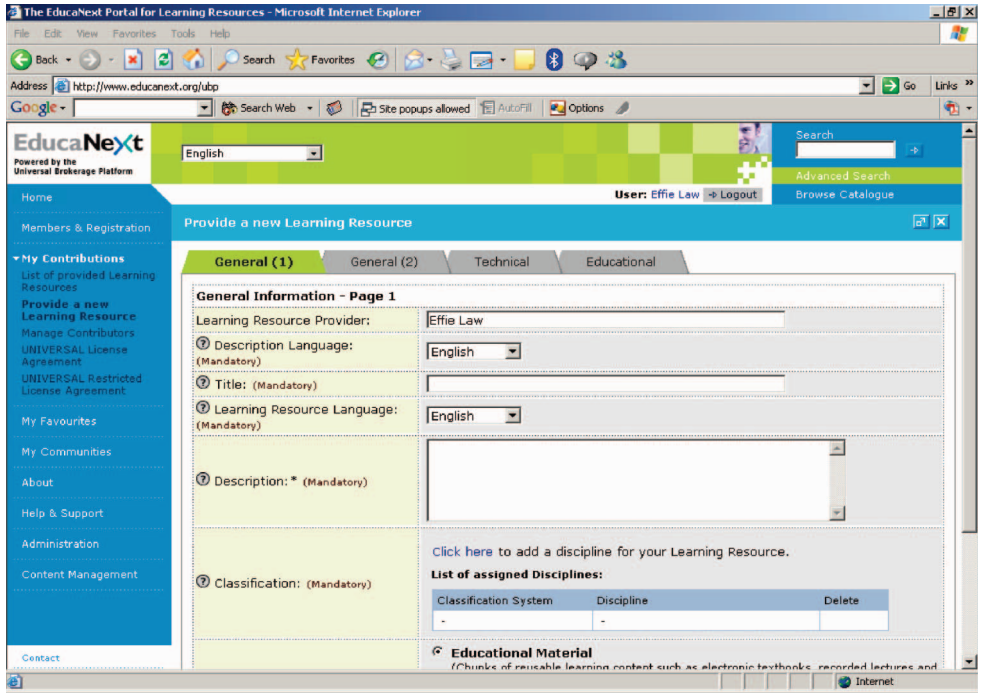

Figure 2. The academic portal usability tested—v. 1.0.

Table 3. Coding scheme of action segments.

Type	Description
Exploratory	Scan the objects on the user interface systematically in a way consistent with a sub-goal
Situated	React almost instantly to a situational demand with embodied skills and knowledge
Planned	Act with anticipation about possible outcomes
Defective	Perform an action leading to an undesirable outcome that may not be noticed immediately
Repair	Act with the goal of fixing a problem that hinders the task progress

In addition, for individual users, a list of usability problems (UPs) was extracted from their respective thinking aloud protocols and the local experimenters' notes. Each UP was categorized based on the scheme depicted in table 1.

6. Findings and interpretations

6.1 *Performance metrics*

A set of quantitative and qualitative measurements were obtained through the local experimenters' observations, analyses of the screen capture records, and the questionnaires. After presenting results, the corresponding hypotheses (section 3) will be discussed.

6.1.1 Task Completion Time (TCT).[3] Note that the TCT for Task 4 was calculated by excluding the time spent on completing the extra Step 5 'Scheduling' (section 4.3) to make the comparisons on an equal footing. Both within-UT (i.e. Task 2 vs. Task 4 of UT1/UT2) and between-UT (i.e. Task2-UT1 vs Task2-UT2 and Task4-UT1 vs Task4-UT2) comparisons in terms of the mean TCT were performed (table 4). Specifically, the TCT of Task 2 over 10 participants of UT1 was significantly higher than that of Task 4 ($t = 3.7$; $df = 9$; $p = 0.005$). Similarly, the mean TCT of Task 2 over 12 participants of UT2 was also significantly higher than that of Task 4 ($t = 3.82$; $df = 11$; $p = 0.003$). Combining the two UTs, the statistically significant difference in TCT between the two tasks over 22 participants was even more salient ($t = 5.09$; $df = 21$; $p = 0.000$). Further, the mean TCT-Task2 of UT1 was significantly higher than that of UT2 ($t = 2.46$, $df = 20$, $p = 0.02$), but there was no significant difference in TCT-Task4 between the two UTs.

[3]It is a well-recognized fact that thinking aloud is not a perfect method in the usability research, especially when time measurement is involved (van den Haak *et al.* 2003), because concurrent verbalization and task performance are two processes that can interfere with each other. An elaborated description on this topic, however, will prolong the length of the paper. A caveat is made here that time-on-task (TOT), which equals TCT minus the verbalization time, may be more accurate than TCT. Due to the relatively small amount of verbalization in our cases, we believe that the difference between TOT and TCT will be insignificant.

H1: It was supported. After attempting a task, the participants in both UTs required significantly less time to work on its variant. This observation directly implies that the participants were able to learn from their experiences and indirectly indicate that the system was highly learnable. Furthermore, the participants required significantly less time in completing the same task with the new version of the system than with the old version. This finding indicated that the improvements, which were partly based on the results of the usability tests on the old version (i.e. UT1), were successful. This serves as a reliable and valid means to evaluate the downstream utility of the user tests.

6.1.2 Usability problems identified. For each user, separate lists of usability problems (UPs) for Task 2 and Task 4 were derived from the raw qualitative data. The UPs identified in Task 4 were further broken down into three sub-groups (table 5):

- *additional*—UPs about the system features that were common to the two tasks, but these UPs were *not* identified earlier in Task 2;
- *specific*—UPs about the system features that were unique to Task 4 (e.g. define the schedule of an educational activity);
- *repeated*—UPs were already identified in Task 2 and re-uncovered in Task 4.

To compute the 'real' difference in the number of UPs between the two tasks, it was necessary to first deduct the number of specific UPs from the absolute number of UPs of Task 4. The mean number of UPs identified in Task 2 over 22 participants was higher than that in Task 4 and the difference was statistically significant ($M_{\text{diff}} = 2.18$; $SD_{\text{diff}} = 3.0$; $t = 3.41$; $df = 21$, $p = 0.003$). Furthermore, the mean number of UPs identified in Task 2 of UT1 was significantly higher than that of UT2 ($t = 3.33$; $df = 20$; $p = 0.003$). Similarly, the mean number of UPs identified in Task 4 of UT1 was also significantly higher than that of UT2 ($t = 2.5$; $df = 20$; $p = 0.02$). These findings imply the effectiveness of the changes introduced into the new version of the system.

Table 4. Mean task completion task per task per UT.

TCT (min.)		UT1	UT2	Combined
Task 2	Mean	14.6	10.8	12.6
	(SD)	(3.6)	(3.6)	(4.0)
Task 4	Mean	8.7	7.4	8.0
	(SD)	(3.0)	(3.6)	(3.3)

Table 5. Number of usability problems experienced by the participants in UT1 and UT2.

	UT1 ($n = 10$)	UT2 ($n = 12$)
Task 2	$M = 5.3$, SD = 2.4	$M = 2.3$, SD = 1.8
Task 4	$M = 4.0$, SD = 3.7	$M = 1.4$, SD = 1.1
Additional	$M = 2.1$, SD = 2.7	$M = 0.6$, SD = 0.5
Specific	$M = 1.4$, SD = 1.0	$M = 0.8$, SD = 0.6
Repeated	$M = 0.5$, SD = 0.7	$M = 0.1$, SD = 0.3

H2: It was supported. The number of UPs identified when performing a task was significantly higher than that when performing its variant subsequently. It can be explicated by the fact that the participants have learnt from their mistakes committed earlier and also developed some workarounds to eschew problems. For instance, in attempting the sub-task of defining an evaluation questionnaire for the learning resource provided for Task 2, some participants (e.g. E9, E10) experienced great difficulty in creating a new questionnaire; they were then completely lost and became very frustrated. When performing the same sub-task for Task 4, these participants simply selected one of the given questionnaires that fitted their learning resources. Interestingly, some other participants handled the same sub-task (i.e. E11) just the other way round, i.e. selecting an existing questionnaire (an easy action) for Task 2 and creating a new questionnaire (a challenging action) for Task 4. This behavior is one of the examples for the category 'additional' UPs and is consistent with Suchman's (1987) claim that people tend to view the same situation differently when they re-explore it and thus uncover new problems.

Further, the significantly shortened task completion time (section 6.1.1) and the significantly lower number of usability problems when performing Task 4 than Task 2 can be interpreted as the evidence for successful learning or positive near transfer (section 2.4). On the contrary, the category 'repeated' (table 5) represents cases of failure to learn or transfer, although on average the number of instances was low. In UT1 there were five 'repeated' cases committed by four participants: one of them was severe, two moderate and two minor, and there was only one 'repeated' case (minor) in UT2. When working on Task 2, E3 experienced the severe UP when he clicked the 'Back' button of the browser (despite the warning message shown earlier) and lost all the data inputted. He was extremely frustrated then. However, he repeated the same 'mistake' when he performed Task 4. This case seems to imply that the procedural knowledge the participant gained in the new task could not override his prior experience acquired when working with other interactive systems. In other words, negative transfer took place because of the mismatch between conditions and actions. In another case, E5 repeatedly overlooked the same mandatory field and left it unfilled. Interestingly enough, in both tasks it took him a while to spot the same error message to know what had gone wrong. It seems to imply that his schema about the

elements of the task (i.e. declarative knowledge) could not be constructed effectively in the first instance and no transfer was possible.

6.1.3 Instances of errors and help. Errors counted include menu-choice error, select-from-list error and others. The same strategy of filtering out 'specific' error (e.g. overlooking the button 'Add Appointment'—a function specific to Task 4) and 'specific' help (e.g. looking up the online help-text for 'Time Zone'—a metadata attribute specific to Task 4) was applied here. The mean number of errors of Task 2 ($M = 2.13$, $SD = 2.06$) over all the 22 participants was higher than that of Task 4 ($M = 1.2$, $SD = 1.37$), but the difference was not statistically significant. Instances of help counted include soliciting help from the local experimenters as well as looking up the online help-text in the system. The mean instance of help seeking of Task 2 ($M = 1.47$, $SD = 1.89$) over all the 22 participants was slightly higher than that of Task 4 ($M = 1.2$, $SD = 1.97$).

H3: It was rejected. In fact, when performing Task 4 some users looked up the online help-text for certain metadata attributes, which they did not bother to do so when performing Task 2. One user remarked 'See what it says about 'Format' [a field] ... that's more or less like what I guessed, fine!' In other words, the users counted on their prior knowledge and did not make a small effort to look up the information available even when they were uncertain about the system's attribute or their own knowledge. Further, in performing Task 4, some users chose a wrong menu and were immediately aware of their own mistake and corrected it. Such 'additional' instances of help seeking and 'avoidable' errors may partially explain the insignificant differences.

The users' help-seeking strategies can be understood with Gray and Fu (2004)'s framework on soft constraints in interactive behavior. Accordingly, users of interactive systems tend to rely on imperfect knowledge in-the-head and ignore the perfect knowledge in-the-world, even when the absolute difference in the effort required is small and even when the reliance on memory is likely to cause a higher error rate and lower performance. Gray and Fu (2004) aim to understand how interactive behavior emerges from the constraints and opportunities provided by the interaction of embodied cognition (Wilson 2002) with the task being performed and the interface designed to perform the task. Besides, they hypothesize that cognitive, perceptual and action operators are orchestrated into patterns of interactive behavior. Indeed, their idea is consistent with our assumption on patterns of cognitive activities (section 2.2). Further, we speculate that the users' (mis)perceived time pressure have somehow driven them to complete the task in the shortest possible time rather than with the highest possible accuracy. The decision on weighing the tradeoffs may relate to whether the users would attribute the slow performance (i.e. long task completion time) to their own ability or to the system's. Psychological attribution theory (Weiner 1986) may shed some light into this phenomenon. None the less, the data of our experimental study, in which the variables of interest were not manipulated to address this specific issue, do not allow us to verify our conjectures. Future work on help-seeking strategies of users of interactive systems is called forth (cf. Martin *et al.* 2005).

6.1.4 Perceived ease and perceived efficiency. The first two questions in the 'After Scenario Questionnaire' (Lewis 1995) measured the user's perceived ease and perceived efficiency of completing individual task, using a 7-point Likert scale with left and right anchors of 'Strongly Disagree' and 'Strongly Agree', respectively:

Q1: Overall, I am satisfied with the *ease* of completing the tasks in this scenario
Q2: Overall, I am satisfied with the *amount of time* it took to complete the tasks in this scenario

The mean perceived ease of completing Task 2 over all the 22 participants ($M = 4.37$; $SD = 1.51$) was slightly lower than that of Task 4 ($M = 4.68$; $SD = 1.36$). The perceived efficiency for Task 2 and Task 4 over all the 22 participants was the same with the value of 4.7.

H4: It was rejected. The user perceived ease and efficiency hardly changed with the task. One possible explanation is that the ratings could not be filtered as in the case of the objective measures (section 6.1.1 – 6.1.3). In other words, when the users evaluated Task 4 they also took into account the extra step 'Scheduling', which for some users was error-free and entailed negligible time and effort whereas for some others was somewhat problematic and thus they experienced lower level of satisfaction.

None the less, these subjective measures are not consistent with the objective ones. Indeed, there were no significant correlations between the perceived ease of task completion and the number of UPs identified in Task 2 or Task 4. No significant correlations between the perceived efficiency of task completion and the TCT of Task 2 or Task 4 could be found either. These results corroborate those of the previous studies that subjective ratings and objective measures of performance do not necessarily correlate (Kissel 1995, Nielsen and Levy 1994, Yeo 2001), and interestingly the magnitude of such a correlation varies with the users' level of computer experience. To verify this observation, the 22 participants of the present study were categorized into experienced and inexperienced computer users, based on their self-reported level of experience in operating database systems, using a 5-point Likert scale with 1 being 'very poor' and 5 being 'very rich'. However, no significant correlation between any of the aforementioned

subjective and objective measures could be obtained. Other factors such as level of competence in information and communication technologies (ICT) and experience in e-learning (i.e. the domain of the system usability tested) did not have any effect on the relationship between the two types of measures either. This particular issue can further be addressed under the framework of Technology Acceptance Model (Moody *et al*. 2003).

To further explore the issue on the mixed evidence concerning the correlation of usability criteria (cf. Frøkjær *et al*. 2000), we computed correlations among the objective performance measures, including task-completion-time (TCT), number of unfiltered usability problems (UPs), instances of help (HELP), and number of errors (ERROR) for the two tasks performed by the 22 users (i.e. $n = 44$). As shown in table 6, the four types of measures are significantly correlated, albeit to different extents. Interestingly, when we broke down the data into Task 2 and Task 4, significant correlations between some pairs of variables were found in the case of Task 2 but not in the case of Task 4 (e.g. TCT-ERROR), or vice versa (e.g. UPs-HELP). The significant correlation between TCT and UPs suggests that more UPs could be found if the users spent more time in tackling the given tasks. Alternatively, the finding may imply that the UPs have prolonged the users' working time. While no conclusive explanations can be drawn, it is clear that the correlations between usability criteria, be they subjective or objective, are influenced by moderator variables such as the level of user computer experience as identified in the previous studies and the nature of the task in the present study.

Table 6. Correlations among four objective performance measures.

1st Variable	2nd Variable	Pearson *r*	Sig. level *p* (two-tailed)
TCT	UPs	0.58**	0.00
UPs	HELP	0.47**	0.00
HELP	ERROR	0.52**	0.001
TCT	ERROR	0.38*	<0.05
TCT	HELP	0.33*	=0.05
UPs	ERROR	0.33*	<0.05

6.2 *Patterns of cognitive activities*

As explained in section 5, user actions were segmented and coded. Note that for Task 4 user actions invoked by completing the extra Step 5 'Scheduling' were excluded. The mean number of action segments of Task 2 was significantly higher than that of Task 4 (table 7) ($M_{diff} = 7.36$; $SD_{diff} = 8.44$; $t = 4.09$, $df = 21$, $p = 0.001$). Intuitively, the number of action segments derived from a task is an approximate indicator of the time a user spent in the task. The Pearson correlation between these two parameters was highly significant ($r = 0.733$).

H5: It was partially supported. The patterns of cognitive activities of Task 2 and Task 4 were obviously different. As shown in figure 3, in accomplishing Task 2 about 34% and 41% of the actions were exploratory and situated, respectively, whereas more than 50% of the actions were planned-based when accomplishing Task 4 ($n = 22$). Contrary to our expectation, the percentages of defective and repair actions were higher in Task 4, with the mean absolute numbers of defective actions being 13 and 10 for Task 4 and Task 2, respectively. It may be attributed to the observation that some users tried out different options when performing Task 4, resulting in more perturbations and concomitant repairs.

Table 7. Number of action segments per task.

Segments		UT1	UT2	Combined
Task 2	Mean (SD)	30 (9.4)	26.4 (8.3)	28.1 (7.7)
Task 4	Mean (SD)	23.7 (8.0)	18.2 (8.1)	20.7 (7.1)

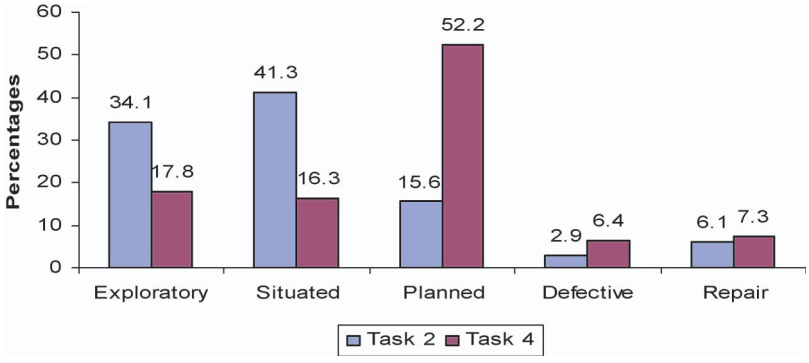

Figure 3. Distribution of 5 types of cognitive activities for Task 2/Task 4 over 2 usability tests.

H6: It was supported. The category of exploratory actions was of particular interest as they represented the users' efforts to validate and adapt their mental models to the particularities of the situation in which they were embedded. The lower the user's level of domain knowledge required for interacting effectively with the system, the higher the tendency that the user explores the situation to bridge the knowledge gap. Indeed, there was a statistically significant negative correlation (Pearson $r = -0.78$; $p = 0.015$; $n = 22$) between the users' self-reported level of expertise (5-point Likert scale; 1 is lowest; $M = 2.8$; $SD = 1.15$) in e-learning (i.e. the domain of the system usability tested) and the percentages of exploratory actions in Task 2.

6.3 *Categorization of usability problems*

Usability problems are defined as problems that hinder users from completing a given task with the system to achieve a specific goal in an effective and efficient manner, or arouse frustration, confusion or some other negative emotion in users when doing so. Further, we classified the UPs into the three categories (cf. section 2.3), namely perceptual motor skills (e.g. the hyperlink for defining the discipline of a learning resource was hardly perceivable), familiar cognitive skills (e.g. the user filled in all four search criteria to locate an existing contributor but actually filling in one criterion was enough; frustration was aroused because the user found it tedious to fill in too many fields), and problem-solving skills (e.g. the conceptual difference between 'providing' and 'offering' learning resources). The mean numbers of UPs of each category for Task 2 and Task 4 are presented in table 8.

Apparently, more usability problems associated with the problem-solving skills level were identified in Task 2 than in Task 4. The difference was statistically significant ($t = 2.29$, $df = 21$, $p = 0.03$) and could be attributed to the fact that the users' conceptual knowledge was relatively deficient when accomplishing Task 2. This finding was also consistent with the observation that the users generally showed higher percentages of exploratory actions in Task 2 than in Task 4 (section 6.2). Nevertheless, the differences in the numbers of UPs associated with perceptual motor and familiar cognitive skills between the two tasks were not significant.

Table 8. Distribution of different types of usability problems.

	Perceptual-motor	Familiar cognitive	Problem-solving
Task 2	1.14	1.0	1.50
Task 4	0.78	1.1	0.68
Additional	0.41	0.6	0.3
Repeated	0.21	0.0	0.0
Specific	0.16	0.5	0.38

Similarly, we categorized the UPs of Task 4 into three subgroups (cf. section 6.1.2). It was interesting to observe that the incidence of 'Repeated' UPs was relatively low and all of them were associated with perceptual motor skills. In fact, when performing Task 4 only four of the 22 users experienced one or two UPs that had already been known in Task 2. It implied that the other users were able to work around some of the UPs they had experienced earlier. Besides, the number of UPs associated with problem-solving skills was the least for the 'Additional' UPs, whereas it was the highest for the 'Specific' UPs. The former could be explained by the fact that the users could somehow clarify some misconceptions about the system features *after* performing Task 2, whereas the latter were related to the new system features (unique for Task 4) for which the users had not yet adapted their mental models.

H7: It was supported. Nevertheless, a particular type of skill—*prototype-using* skills (Bainbridge 1997)—relevant to the current analysis has not yet been addressed. People employ prototype-using skills when they respond to a situation by referring to what is done in a typical situation of the same general type. Put briefly, this skill entails analogical reasoning or transfer of knowledge (section 2.4; section 6.1.2). When performing Task 2, the users might draw on their memories of dealing with a similar system. Note that positive or negative transfer may occur, depending on the compatibility between the experiences drawn upon from working with other systems and those required by the current own. Nevertheless, we cannot identify any evidence in the users' thinking aloud protocols whether and when they have used prototype-using skills when tackling the task on the first instance.

7. General discussions

7.1 *User rational action model*

Generally speaking, the above findings are consistent with the assumption that users behave rationally when working with an interactive system. First, the ability to reflect on one's cognitive state (meta-cognition) is a prerequisite for rational actions. The users' behaviors indicate that they were aware of the fact that their own knowledge state somewhat deviated from the optimal level required for interacting effectively with the system. Consequently, they engaged in systematic exploratory actions, which were driven by the 'information scent' (Card *et al.* 2001) of individual interface objects, thereby incidentally expanding the solution space and improving their mental models. Second, the users demonstrated their capability of learning from performing a task variant, as evident by the significantly lower incidence of exploratory actions and the significantly lower instances of usability problems that were associated with the user's problem-solving skills. Apparently,

the users were able to reason for the knowledge-based usability problems that they had initially experienced when performing Task 2, and to develop some workarounds to avoid the same problems when performing Task 4. This ability to learn and reason is also requisite for rational actions. Third, with the two tasks sharing similar goals and similar contexts (i.e. web-pages), the variation of the user behaviour was primarily determined by the user's knowledge state. When performing Task 4 the way the users interacted with the system features that were common in Task 2 was highly predictable. Indeed, predictability (or consistency) is another key characteristic of rational actions.

Interestingly, the usability problems associated with processing signals (i.e. perceptual motor skills) were somewhat persistent, though to a relatively low extent, as shown by the subgroup 'Repeated' in Task 4. Besides, some more usability problems of this type were identified in Task 4, as indicated by the subgroup 'Additional' (table 8). This phenomenon can be explained by the assumption that the users were preoccupied with the more cognitively demanding task of adapting their mental models to the specific features of the system and thus relegated the saliency of perceptual cues. Otherwise, they would be cognitively overloaded. With the improved mental models, the users were able to navigate the system more effectively when performing the task variant and then became more attentive to the cues displayed in the system. Indeed, a large portion of the additional usability problems identified in Task 4 were related to inappropriate and inconsistent presentation of icons, buttons, links, and feedback.

In summary, we illustrate the foregoing analyses with User Rational Action Model[4] (figure 4). First, users attempt to reach the (sub-)goals of the given tasks by satisfying situational demands. Next, users check the compatibility between the demands and their own knowledge level. If the two parameters converge, then plan-based actions will be executed. Otherwise, exploratory and situated actions will be performed to overcome the 'representational bottleneck' (Wilson 2002). Both situated and plan-based actions will change the system state and the resulting feedback can somehow update the users' mental models about the system. Iteratively, the users check the status of goal attainment and repeat the cycle, if required.

[4]The TOTE (Test-Operate-Test-Exit) model (Miller *et al.* 1960) is fundamental to cognitive psychology and the information processing framework. Hence, there is no doubt that our User Rational Model, just like the contemporary work on user modeling in HCI, is related to TOTE. However, we move beyond TOTE and other similar work (e.g. Blandford *et al.* 2001) to address user action by incorporating the concept of situated cognition, given the understanding that human behaviour is not entirely plan-based (Suchman 1987). Further, we embrace the hybridized model that amalgamates the cognitivist-rationalist and situated-constructivist paradigms (section 1), and the User Rational Action Model can well summarize our understanding in this regard.

7.2 *User-based evaluation tests*

User-based evaluation has extensively been applied in industry because of trustworthy results it normally yields. However, this approach has a number of drawbacks. One of which is to adequately train users to manage advanced functions of a relatively complex interactive system. Indeed, complexity is one of the various dimensions for classifying software systems; it is a continuum rather than a dichotomy. The use of a system of one end of this continuum requires minimal learning or experience (e.g. Google), whereas the use of a system of the other end may entail systematic training (e.g. aviation traffic control). The system we have evaluated is somewhere in between. Specifically, the tasks enabled by this system involve a sequence of steps and a number of options; users do need some practical experience in order to work with it effectively. We argue that including task variants in user tests for evaluating moderately complex systems such as the one examined in this study is advantageous, based on the following considerations.

- *Increase the validity of the evaluation results*. It addresses the issue of 'at least two instances'. Some usability practitioners tend to discard those usability problems that are reported only once (Lewis 1994). Our findings suggest that quite a number of 'usability problems' were actually related to the users' deficient mental model. If the system tested is highly learnable, some usability problems identified in the initial interaction with the system will not appear in subsequent interactions. Instances of 'False Alarm' will thus be reduced. Developers can then focus on fixing those persistent problems.
- *Increase the thoroughness of the evaluation results*. In performing a task variant, users may attempt options that they have not yet tried out earlier and thus identify more usability problems. In this sense, the absolute number of users required to identify a certain percent of usability problems can be reduced.
- *Provide insights into the re-design of the system evaluated*. Observing how users work around the difficulties that they have experienced. These work-around strategies can serve as improvement suggestions for developers.
- *Provide viable and valid measures of learnability*. Learnability can be measured by comparing a novice user's initial and improved performance that is enabled by a period of training. For a system of moderate complexity that enables learning-by-doing-with-minimal-instruction, performing a task and its variant can be an effective means of advancing a novice user's knowledge and skills. The learnability of the system can be attested with a high level of confidence if positive outcomes are obtained especially

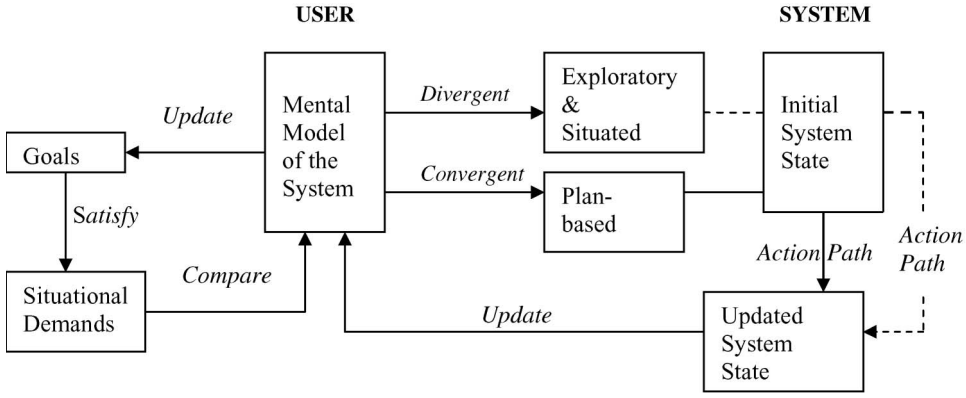

Figure 4. User rational action model.

when no extra training is required and one-shot practice trial is proved adequate for a novice user to achieve experienced user performance (Moyes and Jordan 1993). As an alternative to the traditional way of measuring learning time or effort, the differences in the number and nature of usability problems between performing a task and its variant can be a valid measure of learnability as well.

Put briefly, the overall effectiveness of user-based evaluation will be enhanced when task variants are included. However, performing task variants implies prolongation of a test session. It may not only increase the costs of running the test but also lower the user motivation. Such disadvantages may become more acute when a system and its component tasks are very complex. Hence, using task variants may be more cost-effective for moderately complex systems. Indeed, the problem of task selection (Cordes 2001) for usability evaluation is a tricky issue that calls forth more research efforts.

Furthermore, given that carrying out a task variant can significantly improve user performance in a relatively complex task, the practical implication is how a novice user can be enticed to attempt the same task again before leaving the system. Presumably, such a second attempt may enhance the casual user's retention of the system use and thus re-learnability. It may also improve the user's satisfaction through mastering the complex task and thus her overall acceptance towards the system. One possible means to encourage the user to make a second attempt is to present a projected learning curve (i.e. estimated time and effort to achieve experienced user performance) based on her own initial performance and on the cumulative data of the other users' performances.

8. Concluding remarks

Using task variants is a common technique employed for reinforcing learning in everyday educational context and for checking the reliability of behavioural performance of interest in psychological experiments. However, using task variants in user-based usability evaluation is relatively uncommon. One reason might be that evaluators do not know or appreciate the value of using task variants. Another reason might be that usability evaluations vary widely, and that detailed evaluations of the type presented in this paper are not always seen as 'cost-justified'. As an applied field usability evaluation is inevitably pragmatic (i.e. being guided by practical experience and observation rather than theory) and driven by cost-effectiveness. None the less, without understanding 'why' and 'how' a usability evaluation method works, there is always a risk that evaluators may choose a wrong technique or tool, and the loss thus incurred could be great. The literature shows that there have not been any meticulous analyses of the effects and underlying cognitive mechanisms of using task variants as presented in this paper. Indeed, there is a lack of such analyses that are imperative for substantiating the intellectual depth of the field of usability as a whole.

Results showed that generally the users could derive strategies for working out task variants and the system was proved highly learnable. The methodologies presented in this paper for assessing the rationality and adaptability of users and for evaluating the learnability of an interactive system are practical. However, they need to be further validated and improved with an even larger sample of users and different designs of task variants. One significant topic that the current paper does not address is error recovery. In our tests, most of the users tended to repeat the same sequence of actions when impasse arose. Besides, we observed several instances of the so-called 'garden path situation' (Suchman 1987) that the users failed to identify some human–machine communicative trouble at the point where it occurred, and discovered only at some later point in the interaction, but it was then too difficult to find out the source of the trouble. It is intriguing to find out how users become aware that they have taken the wrong path in the menu hierarchy and how they decide to undertake some

corrective actions (Curzon and Blandford 2000). It is also important to know how users learn from their problems in interacting with the system and how effective their work-arounds are. Another topic worthy of further exploration is how subjective measurements are related to objective measurements and which moderator variables significantly influence the relationship. The inconclusive findings hitherto garnered suggest that the issue remains open. Further, it is intriguing to deepen our understanding about the transfer of task knowledge by systematically controlling the order in which target tasks are presented, the number of intervening tasks between the target tasks, and the degree to which the target tasks are similar to each other. Clearly, all the foregoing challenging research questions entail more systematic empirical studies.

Acknowledgements

We are very grateful to the insightful comments given by the three anonymous reviewers on the draft of this paper. Thanks should also go to the test participants who voluntarily took part in the usability evaluation tests. We would like to express our gratitude to the two projects: COST Action 294 (http://www.cost294.org) and PROLEARN (http://www.prolearn-project.org), which enabled the collaboration between the three authors.

References

Anderson, J.R., 1993, *Rules of the Mind* (Hillsdale, NJ: Erlbaum).

Bainbridge, L., 1997, Multiplexed VDT display systems. In *Human Computer Interaction and Complex Systems*, edited by G.R.S. Weir and J.L. Alty, pp. 181–210 (New York: Academic Press).

Blandford, A., Butterworth, R. and Curzon, P., 2001, PUMA Footprints: linking theory and craft skill in usability evaluation. In *Proceedings of INTERACT '01*, 9–13 July 2001, Tokyo, Japan, edited by M. Hirose, pp. 577–584 (Amsterdam: IOS Press).

Card, S.K., Moran, T.P. and Newell, A., 1983, *The Psychology of Human Computer Interaction* (Hillsdale, NJ: Lawrence Erlbaum Associate).

Card, S.K., Pirolli, P., van Der Wege, M., Morrison, J.B., Reeder, R.W., Schraedley, P.K. and Boshart, J., 2001, Information scent as a driver of Web Behavior Graphs. In *Proceedings of CHI 2001*, Seattle, WA, USA.

Cordes, R.E., 2001, Task-selection bias. *International Journal of Human–Computer Interaction*, **13**(4), 411–419.

Curzon, P. and Blandford, A., 2000, Reasoning about order errors in interaction. In *Supplementary Proceedings of the 13th International Conference on Theorem Proving in Higher Order Logics*, August 2000, Portland, US, pp. 33–48 (Portland, USA: Oregon Graduate Institute).

Dumas, J.S. and Redish, J.C., 1999, *A Practical Guide to Usability Testing* (revised edition) (Bristol, UK: Intellect).

Eason, K., 1984, Towards the experimental study of usability. *Behaviour and Information Technology*, **3**(2), 133–143.

Frøkjær, E., Hertzum, M. and Hornbæk, K., 2000, Measuring usability: are effectiveness, efficiency and satisfaction really correlated? In *Proceedings of CHI 2000*, 1–6, April, the Hague, Netherlands, pp. 345–352 (New York: ACM).

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

Gillan, D.J. and Bias, R.G., 2001, Usability science. 1: Foundations. *International Journal of Human–Computer Interaction*, **13**(4), 351–372.

Gray, W.D. and Fu, W-F., 2004, Soft constraints in interactive behaviour. *Cognitive Science*, **28,** 359–382.

Greeno, J.G., Collins, A. and Resnick, L.B., 1996, Cognition and learning. In *Handbook of Educational Psychology*, edited by D.C. Berliner and R.C. Calfee, pp. 15–46 (New York: Macmillan).

Gruber, H., Law, L.-C., Mandl, H. and Renkl, A., 1995, Situated learning and transfer: state of the art. In *Learning in Humans and Machines. Towards an Interdisciplinary Learning Science*, edited by P. Reimann and H. Spada, pp. 168–188 (Oxford: Pergamon).

Johnson, T.R., Wang, H. and Zhang, J., 1998, Modeling speed-up and transfer of declarative and procedural knowledge. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, pp. 531–536 (Hillsdale, NJ: Erlbaum).

Karasavvidis, I., Pieters, J.M. and Plomp, T., 2000, Investigating how secondary school students learn to solve correlational problems. *Learning and Instruction*, **10,** 267–292.

Kirakowski, J. and Corbett, M., 1988, Measuring user satisfaction. In *People and Computers IV*, edited by D.M. Jones and R. Winder, pp. 309–328 (Cambridge: Cambridge University Press).

Kissel, G.V., 1995, The effect of computer experience on subjective and objective software usability measures. In *Proceedings of CHI' 95*, pp. 284–285 (New York: ACM Press).

Lemon, B., Pynadath, D., Taylor, G. and Wray, B., 1994, Cognitive architectures. Available online at: http://ai.eecs.umich.edu/cogarch4/index.html

Lewis, J.R., 1994, Sample sizes for usability studies: additional considerations. *Human Factors*, **36**(2), 368–378.

Lewis, J.R., 1995, IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *International Journal of Human–Computer Interaction*, **7**(1), 57–78.

Martin, A., Ivory, M., Slabosky, B. and Megrew, R., 2005, How helpful is help? Use of and satisfaction with user assistance. In *Proceedings of UAHCI 2005 (the 3rd International Conference on Universal Access in Human Computer Interaction)*, 24–27 July 2005, Las Vegas, USA.

Miller, G.A., Galanter, E. and Pribram, K. 1960, *Plans and the Structure of Behaviour* (New York: Holt).

Moody, D.L., Sindre, G., Brasethvik, T. and Sølvberg, A., 2003, Evaluating the quality of information models. In *Proceedings of International Conference Software Engineering (ICSE) 2003*, 3–10 May 2003, Portland, Oregon, pp. 295–307 (IEEE Computer Society).

Moyes, J. and Jordan, P.W., 1993, Icon design and its effect on guessability, learnability, and experienced user performance. In *People and Computers VIII*, edited by J.L. Alty, D. Diaper and S. Guest, pp. 49–59 (Cambridge: Cambridge University Press).

Newell, A., 1990, *Unified Theories of Cognition* (Cambridge, MA: Harvard University Press).

Newell, A. and Simon, H.A., 1976, Computer science as empirical inquiry: symbols and search (1975 ACM Turing Award Lecture). *Communications of ACM*, **19**(3), 113–126.

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

Nielsen, J. and Levy, J., 1994, Measuring usability: preference vs. performance. *Communications of the ACM*, **37**(4) 66–75.

Perkins, D. and Salomon, G., 1994, Transfer of learning. In *International Encyclopaedia of Education*, pp. 6452–6457 (Oxford: Elsevier).

Polson, P.G. and Kieras, D.E., 1985, A quantitative model of the learning and performance of text editing knowledge. In *Proceedings of CHI'85*, 14–18 April 1985, San Francisco, edited by L. Borman and B. Curtis, pp. 207–212 (New York: ACM).

Rasmussen, J., 1986, *Information Processing and Human–Machine Interaction: An Approach to Cognitive Engineering* (Holland: Elsevier).

Rosson, B. and Carroll, J., 1995, Narrowing the specification-implementation gap in scenario-based design. In *Scenario-based Design: Envisioning Work and Technology in System Development*, edited by J.M. Carroll, pp. 247–278 (New York: Wiley).

Shackel, B., 1986, Ergonomics in design for usability. In *People & Computers: Designing for Usability*, edited by M.D. Harrison and A.F. Monk, pp. 44–64 (Cambridge: Cambridge University Press).

Stoke, D.E., 1997, *Pasteur's Quadrant: Basic Science and Technological Innovation* (Washington, DC: Brookings).

Suchman, L., 1987, *Plans and Situated Actions* (Cambridge: Cambridge University Press).

van den Haak, M.J., de Jong, M.D.T. and Schellens, P.J., 2003, Retrospective vs concurrent think-aloud protocols: testing the usability of an online library catalogue. *Behaviour & Information Technology*, **22**(5), 339–351.

Weiner, B., 1986, *An Attributional Theory of Motivation and Emotion* (New York: Springer-Verlag).

Wilson, M., 2002, Six views of embodied cognition. *Psychonomic Bulletin & Review*, **9**(4), 625–636.

Yeo, A.W., 2001, Global-software development lifecycle: an exploratory study. In *Proceedings of CHI'01*, pp. 104–111 (New York: ACM Press).

Appendix A. An example of action segment analysis of Task 2 performed by E2

Table A1. Categorization of action segments.

Action segment	Categorization
Look at the lists of Educational Materials and Educational Activities on the homepage	EXPLORATORY—improve the mental model of the system
Scan the items of each of the main menus on the homepage: My Contributions → My Booking → About EducaNext → Help & Support	EXPLORATORY—identify appropriate path and improve mental model about the system
Click 'Help & Support' → How to get support → FAQ	EXPLORATORY—improve the mental model of the system
Click the menu 'My Contributions' → list of provided Learning Resources	EXPLORATORY—expand the solution space; information sent
Login	REACTIVE—embodied practice
Scan the menu items: 'Details' → 'Booking Statistics'	EXPLORATORY—improve the mental model of the system
Scan the menu bar on the left-hand side column	EXPLORATORY—improve mental model and evaluate the scope of solution path
Click the sub-menu 'Provide a New Learning Resource'	PLANNED—align with the objective described in the task instruction
Fill in the attribute: Description	REACTIVE
Fill in the attribute: Title	REACTIVE
Fill in the attribute: Type	REACTIVE
Click the blank box of the field 'Discipline' and leave it empty	DEFECTIVE—garden path (no warning from the system; the user was unaware of the error)
Click 'Next'	REACTIVE—affordance
Fill in the attribute: Contributors/Authors	REACTIVE
Upload an Learning Resource	PLANNED—aligned with the objective given in the task instruction
Fill in the technical information	REACTIVE
Click 'Finish'; Bounced back to Step 1, overlook the error message	REACTIVE
Click 'My Contribution'	REPAIR—backward tracking
Click 'Back' of the browser → Step 1 of providing new Educational Material	REPAIR—backward tracking
Click 'Next' → warning message from the system → front page of 'My Contribution'	REPAIR—locate causes
Scan the menu bar on the left-hand side column and the buttons on the top toolbar	REPAIR—an evaluative action for locating causes
Click 'Shared selected Learning Resource'	REPAIR—miscued by the label
Click 'Provide a New Learning Resource' (start from scratch again)	PLANNED—revisit
Re-enter data for 'Title' and 'Description'	PLANNED—revisit

(*continued*)

Table A1. (*Continued*).

Action segment	Categorization
Click help-text for 'Classification'	EXPLORATORY—enhance mental model
Click the blank box of the field 'Discipline' and leave it empty	DEFECTIVE
Click 'Next' → author/contributors → upload →	PLANNED—revisit
Click 'Finish' → bounced back to step 1 with error message	PLANNED—repeat
Click the hyperlink 'Click here' and select a value for 'Discipline'	REPAIR
Click 'yes' for Offer	REACTIVE
Choose 'UNIVERSAL License Agreement'	REACTIVE
Click 'save offer'	REACTIVE