Towards a Framework for Usability Testing of Interactive e-Learning Applications in Cognitive Domains, Illustrated by a Case Study

S.S. (THABO) MASEMOLA AND M.R. (RUTH) DE VILLIERS
University of South Africa

Testing has been conducted in a controlled usability laboratory on an interactive e-learning application that teaches mathematical skills in a cognitive domain. The study obtained performance measures and identified usability problems, but was focused primarily on using the testing technology to investigate distinguishing aspects of such applications, such as time usage patterns in domains where rapid completion is not necessarily a performance indicator. The paper addresses the issue of what, actually, is meant by 'usability' in learning environments. A pilot study identified obstacles and served to enhance the main study. Thinking-aloud on the part of participants provided useful data to support analysis of the performance measures, as can benchmarks and best case measures. Particular attention must be paid to ethical aspects. Features emerging from this study can contribute to a framework for usability testing and usage pattern analysis of interactive e-learning applications in cognitive domains.

Categories and Subject Descriptors: H.5.1 [**Multimedia Information Systems**] *Evaluation/methodology*;
H.5.2 **[User Interfaces]** *Benchmarking, Evaluation/methodology, Interaction styles, screen design*;
K.3.1 [**Computer use in Education**] *Computer-assisted learning*; K4.1 **[Public Policy Issues**] *Ethics*
General Terms: Design, Experimentation, Human Factors, Measurement, Performance
Additional Key Words and Phrases: Cognition, e-learning, human-computer interaction, reflection, think-aloud, usability testing.

1. INTRODUCTION

Usability testing is a software evaluation technique that involves measuring the performance of typical end-users as they undertake a defined set of tasks on the system being investigated. It commenced in the early 1980s, as human-factors professionals studied subjects using interfaces under real-world or controlled conditions and collected data on problems that arose ('human factors' is an early term for the human-computer interaction discipline). It has been shown to be an effective method that rapidly identifies problems and weaknesses, and is particularly used to improve the usability of products [Dumas, 2003; Dumas and Redish, 1999; Jeffries et al, 1991]. Since the early to mid-1990s, such testing has been empirically conducted in specialized controlled environments called *usability laboratories*, equipped with sophisticated monitoring and recording facilities for *formal usability testing*, supported by analytical software tools. It is an expensive technique. Participants, who are real end-users, interact with the product, performing specified representative tasks. Their actions can be rigorously monitored and recorded in various ways: by videotape − for subsequent re-viewing; event logging − down to keystroke level; and audio − to note verbalization and expressions. The data, in both quantitative and qualitative forms, is analysed and changes can be recommended. Typical usability metrics include the time taken to complete a task, degree of completion, number of errors, time lost by errors, time to recover from an error, number of subjects who successfully completed a task, and so on [Avouris, 2001; Dix, Finlay, Abowd & Beale, 2004; ISO 9241, 1997; Preece, Rogers & Sharp, 2003; Wesson, 2002]. The primary targets of usability testing are the user interface and other interactive aspects. Such testing is used by academics for research and development, and also by usability practitioners in the corporate environment for rapid refinement of interfaces and analysis of system usability. In the latter case the outcome is a report documenting suggested improvements, as opposed to validation or rejection of hypotheses [Shneiderman, 1998]. In commercial environments, the process may be done before release, to repair flaws upfront.

The aim of this study is not merely the evaluation of a target system for its own sake, since the application tested has already been investigated by other methods, including the highly effective method of heuristic evaluation. Instead, we used a target system in the process of establishing a framework for formal usability testing, not of business systems, but of interactive e-learning applications in the *cognitive domains* of computational disciplines. Such applications do not involve rote learning, and they differ from drill-and-practice systems for mastery learning. Rather, they entail critical thinking and rigorous reasoning on the third- (analysis), fourth- (synthesis) and fifth- (evaluation) levels of Bloom's [1956] Taxonomy. Our findings suggest innovative ways of using the facilities in usability laboratories.

First, we briefly explain the view of e-learning that underlies this paper and then discuss usability testing of e-learning. Some definitions of e-learning equate it solely with use of the Internet in instruction and learning, but others [CEDEFOP, 2002; Wesson & Cowley, 2003] are broader, including multiple formats and methodologies such as the Internet and Web-based learning (WBL), multimedia CD-ROM, online instruction, educational software/courseware,

Author Addresses:
S.S. Masemola, School of Computing, University of South Africa, P O Box 392, UNISA, 0003, South Africa; masemss@unisa.ac.za.
M.R. de Villiers, School of Computing, University of South Africa, P O Box 392, UNISA, 0003, South Africa; dvillmr@unisa.ac.za.

and traditional computer-assisted learning (CAL). This approach suits the present study, which views e-learning as a broad range of learning technologies encompassing various roles for technology, including interactive educational software, web-based learning, learning management systems, and learners using computers as tools [De Villiers, 2005].

Formal usability testing can enhance the evaluation of e-learning. Although much general research on evaluation has been done in the human-computer interaction (HCI) discipline, there is scope for evaluation and metrics specifically for interactive educational applications [Wesson, 2002]. Van Greunen and Wesson [2002] of Nelson Mandela Metropolitan University, the first South African University to acquire usability lab facilities, propose a methodology for formal usability testing of interactive educational software. They call for more such case studies, but of a different nature from their initial studies, so as to investigate testing on a variety of e-learning products.

This paper builds on these needs, as it describes a case study on usability testing of an e-learning application in a mathematical domain that demands analytical cognitive skills. With the advent of its own usability lab, our home base, the School of Computing (SoC) at UNISA, has new opportunities. Since 1992 the Centre for Software Engineering (CENSE), within the SoC, has designed, developed and evaluated CAL and WBL products for the School and other UNISA departments. This new usability testing technology offers enrichment and potential synergy:

- Quantitative measurement of usability and performance can be added to the existing evaluation methods;
- Frameworks can be set up for usability testing of different kinds of applications.

2. BACKGROUND: THE ENVIRONMENT, THE APPLICATION, AND THE APPROACH

2.1 The environment

The SoC's new usability lab at the Muckleneuk Campus in Pretoria has video and audio monitoring equipment, as well as event logging capabilities. The *Noldus* software has sophisticated analytical tools and reporting facilities. The administrator or researcher managing the session and observing participants is invisible behind sound-proof one-way glass. Communication between the two sections of the locale is via microphone. After the testing, questionnaires or interviews can be used to determine user satisfaction with the product evaluated.

2.2 The application

The application used as a testing target is a CAI lesson in discrete mathematics, called *Relations*. It was developed in the mid-1990s for a first-level module in Theoretical Computer Science, and was used until 2004 when a completely re-designed Version 2 was released. *Relations* involves critical thinking and analytical skills and is optional supplementary educational material (available at a reasonable price). It conforms to Alessi and Trollip's [2001] definition of a CAI *tutorial*, which 'presents information and guides the student'. It is highly interactive, as it presents theoretical concepts, examples, exercises for practice, and hyperlinks to definitions. Teaching is interspersed with practice opportunities in the form of question segments. There are no multiple-choice questions, but a few Yes/No's. Fill-in-the-blank questions often require composite answers or a series of mathematical characters. Some exercises entail synthesis of a relation to meet particular conditions, where more than one alternative can be correct. Detailed diagnostic feedback and explanations are provided, and second attempts must be made after wrong answers. *Relations*gives users the option to omit theory and go straight to exercises or revision screens, thus catering for different stages of learning. It is a multi-media production that runs on a self-contained CD but is not Web-based.

2.3 The approach

Relations Version 2 was developed by a participative action research approach and has been evaluated by triangulated techniques, namely, heuristic evaluation, questionnaire surveys among users, interviews, and a post-test [De Villiers, 2004; 2006]. It has been corrected and refined longitudinally. The idea in this study was indeed to see what further data emerged about *Relations* and what issues were identified by formal usability testing in a controlled environment. But, primarily, the testing was a means to an end in the effort to establish a framework for usability testing that addresses the *unique aspects of testing interactive learning applications in cognitive domains.* Such a methodology will be different from one for testing task-oriented applications, and also different from one for e-training in rote procedures or motor skills. Before discussing how it differs, we note that the general definition of usability refers to effectiveness, efficiency and satisfaction [ISO 9241, 1997]. However, conventional usability testing is not the optimal way to judge applications that support learning which, by its nature, is focused more on a process than on generating a product:

- Efficiency cannot be judged by low times taken on tasks. Learning how to reason with content − in mathematical and scientific disciplines − does not necessarily require rapid progress through content or learning tasks. Users have different learning styles and different approaches to working through courseware or doing exercises. Speed of learning is less a measure of system efficiency than it is a function of personal ability and learning style. A rapid completion rate is not necessarily a 'good' measurement, since the application should foster personalized learning.
- It is not always desirable to minimize errors. Squires and Preece [1999] distinguish between peripheral cognitive errors (usability errors) and true cognitive errors, which are part of the learning process, particularly in complex domains. Usability-related errors should be avoided, but cognitive errors should be permitted, provided that

support mechanisms exist to promote a recognition–diagnosis–recovery cycle. The value of such mistakes is substantiated by Mayes and Fowler [1999:485] who stress that in educational applications a 'seamless fluency of use is not necessarily conducive to deep learning … the software must make learners think'.

– There is a distinction between evaluating functionality and evaluating usability, but in cognitively-oriented applications this horizon blurs. The functional operations undertaken by users are learning activities, so the learning process is part of the instructional functionality. As well as being a computing application, an e-learning product is also course material. The effectiveness of learning and the users' subjective satisfaction with a resource are therefore part of its usability. *So the learning content and the learning process can be viewed as being related both to functionality and to usability*, confirmed by the definition of usability of educational software [Geisert & Futrell, 1995, cited in Van Greunen & Wesson, 2002; Wesson, 2002], which refers to the extent to which it can be used to achieve specified learning outcomes with effectiveness, efficiency and satisfaction in a specified learning context. Similarly, Quintana et al [2003] believe that in evaluating educational software applications, they should also be evaluated to test how well they support learners in learning. Usability in e-learning involves the content itself, as well as the way in which the content is presented [Wesson & Cowley, 2003].

So what are we undertaking? With the object of diagnosing problems, we measure conventional factors such as ease of use by measuring the time spent on a task and the number of errors. But as explained, the aspect of time has associated complexities. The *time taken in accessing* a specified task is a measure of usability, but the *time spent interactively learning* new mathematical concepts is not a usability measure. It is a cognitive usage measure, indicating the participant's learning style, aptitude or cognition pattern. We need to distinguish between these two ways of using time, and can do so by supplementing quantitative technological metrics with the qualitative approach of *thinking-aloud* protocols [Boren & Ramey, 2000; Dumas, 2003]. Regarding functionality and usability, we work on the premise that usability includes the effective attainment of learning, as discussed above. As well as investigating navigation through the interfaces, we are assigning learning tasks to the participants as part of the formal testing. Although time spent on cognitive activities will not be used as a usability metric, it can be considered a *usage metric*. However, performance on specified learning tasks can be measured as part of the *effectiveness* aspect of usability, because successful cognitive processing is intrinsic to both functionality and usability.

3. RESEARCH DESIGN, METHODOLOGY AND TEST PLAN

3.1 Research questions

With the argument in Section 2.3 as a background, the following research questions emerge:

1. In addition to the identification of performance measures and problems, how can testing in a usability laboratory elicit valuable information about cognitive e-learning applications?
2. What activities and outputs yield meaningful information about interactive applications in such domains?
3. What notable features emerge from this study that can contribute specifically to a framework for usability testing of interactive e-learning in cognitive domains?

3.2 Design and Methodology

Research design is based on case study research [Hays, 2004], which involves close examination of people, topics, issues or programs. It aims to answer focused questions by means of in-depth descriptions and interpretations over a fairly short period, uncovering new or unusual interactions, events and explanations. Case studies serve to examine unique aspects in each case. We conducted two cases of usability testing to observe and investigate details of the interaction between humans and an e-learning application. The research design is also an emergent design, in that the experience and findings of the pilot study strongly influenced the approach to the main study.

Dumas [2003] lists six defining characteristics of usability tests, while a seventh and eighth are obtained from Dumas and Redish [1999]:

1. The focus is usability.
2. Participants are end users or potential end users.
3. There is an artifact to evaluate, which may be a product design, a system or a prototype.
4. The participants think aloud as they progress through tasks.
5. Data is recorded and analysed.
6. The results are communicated to appropriate audiences (often a corporate client).
7. Testing should cover only a few features of the product, incorporated in selected tasks.
8. Each participant should spend approximately an hour doing the stated tasks.

Our methodology and test plan are based on general methodologies for formal usability testing [Pretorius, Calitz & Van Greunen, 2005; Rubin, 1994; Van Greunen & Wesson, 2002] but with some distinguishing features, such as the emphasis on participants thinking aloud and the use of a best case for initial benchmarking, to be described in Section 5.2. The broad methodology involves the following steps:

- Set up objectives in line with research questions.
- Determine the aspects to be measured and their metrics.
- Formulate documents:
  Initial test plan, task list, information document for participants, checklist for administrator, and determine a means of investigating satisfaction.
- Acquire representative participants.
- Conduct a pilot test.
- Refine the test plan, task list, and information document for the main usability test in the light of the pilot.
- Conduct usability test.
- Determine means of analysis and presentation that address the unique, as well as the usual, aspects.
- Draw conclusions and make proposals for the way forward.

3.3 Test plan

Planning should be well-managed with a written test plan. The test plan is not explicitly set out here, since it is unpacked in the course of the paper. Moreover, it evolved considerably from the pilot to the main study, as shown in Section 5.2. However, a few notable aspects are mentioned below:

Objectives: As already stated, previous evaluations have been conducted on *Relations*, De Villiers [2004; 2006], but the objective of this study was to add an empirical approach to detecting usability issues and to set initial benchmarks. We are also working towards a general framework for meaningful testing of e-learning applications, involving both usability measures and usage patterns.

Briefing participants and ethics: In usability experimentation, attention to ethical aspects is essential [Dumas, 2003; Rubin, 1994]. A detailed information document was compiled to set out the purpose, procedures and stakeholders in the research. It explained the rights of all parties, detailing anonymity and confidentiality, and pointing out that the sound and video recordings – which violate anonymity – would be used for research purposes only. The document included an informed-consent form.

Thinking aloud: An important aspect is the value we attach to the *think-aloud protocol* over and above the observation [Avouris, 2001; Boren & Ramey, 2000; Dix et al. 2004; Dumas, 2003; Shneiderman, 1998]. Participants verbalize as they work through the required tasks. For example, they describe their thoughts, what they are doing, what they experience, and why they take a particular action. This information tells the observer how they are spending their time during the periods when they are not actively interacting with the application.

4. THE PILOT STUDY

4.1 Process and participants

Since we were novices with respect to usability testing, we conducted a pilot test to try out the methodology, the tasks and measures, and to prepare for technological challenges. Dumas and Redish [1999] describe this as 'debugging' the equipment, software and procedures, as well as providing a practice run. A test administrator requires skills that are acquired with practice, so it was important for the administrator to gain familiarity with the planned activities. The emphasis was not on data analysis, although observations were made about the participants' general experience.

It is not easy to acquire research subjects in a distance-teaching environment like UNISA. By means of a call-for-volunteers in a tutorial letter to all students in *Theoretical Computer Science 1*, we obtained participants for both the pilot study and the main study. The letter explained that not all the volunteers would be participants, but that we would take samples of convenience to reflect some of the variety in the user community. UNISA's student population is highly diverse. A first-year student can be aged between 17 and 50. Some students in Computing have never used a computer before, while others were brought up with computing technology. Some are full-time employed, but an increasing number are young full-time learners in their late teens or twenties.

The optimal number of participants advocated by Nielson [2000] for best use of usability testing resources and optimal identification of problems is five. It can be done with as few as three subjects [Shneiderman, 1998], while Dumas and Redish [1999] advise six to twelve in subgroups with common characteristics. Typically the number of participants is too low for much statistical analysis [Preece et al, 2003]. Fifteen students responded to the call-for-volunteers. We invited five, who compositely represented the heterogeneous population, to participate in the pilot test, while taking care that a similar diverse group was available for the main test. They were asked to work through relevant sections in their study material before attending the testing session. There were four males and a female. Two were full-time employed, professionals with prior qualifications, aged in their thirties. One was a young part-time student and the other two, aged 17 and 20, were full time students. Only the 17-year old had not used a computer before.

4.2 Methods and tasks

The testing investigated some of the effectiveness of Relations as a supplementary learning tool, and time spent on stated tasks, distinguishing between time spent learning the application and its navigation facilities, and time spent on cognitive activities as learners engaged with the content. Usability errors were recorded as errors, but cognitive errors were considered as learning activities. The main aims were for us to understand, first-hand, the complexities of the testing environment, and to gain insight into the specifics of testing e-learning.

The task list gave two major composite tasks, with subtasks. ***Task1*** involved using the main menu to access **Properties of relations** and then to enter the interactive teaching content, visuals and exercises on **reflexivity**, **symmetry** and **transitivity** respectively, these three being the fundamental theoretical concepts of the domain. Then it was back to the main menu and into a submenu for **Special kinds of relations**, an integrated application section, from which participants had to navigate to the section on **Equivalence relations**. ***Task 2*** was simpler. It involved accessing the self-test, **Test Yourself,** and doing specified questions related to the content of Task 1. (Section 5.2 will show how these tasks were restructured and subdivided for the subsequent main study.)

The participants came one at a time by personal appointment to the laboratory, where a two to three hour block was set aside for each session. Although it is a controlled environment, efforts were made to help them relax and simulate as far as possible the situation of using the CAI lesson, *Relations*, at home. They were treated with respect and made to feel partners in the research, rather than passive subjects. The procedure was outlined, explaining that the system was being tested and not them and that 'testing' had no bearing on their module mark. Brief descriptions were given of the think-aloud protocol and what was expected of them, after which they signed the consent form.

4.3 Findings – effectiveness and satisfaction

As stated, the main purpose of the pilot test was to test the process and not to collect and analyse data. However, the video suggests that participants had a positive and satisfying experience. They smiled and made positive sounds, such as 'O...kay…'. All five completed all the tasks. When they made errors, they were able to recover without help from the administrator. Of the five, only one remembered to consistently think aloud while doing the tasks. Another read what was on the screen, but added no personal introspection. The verbalizations of the other three could not be discerned. We decided to interrupt them twice to remind them to think aloud, but thereafter to leave them alone. After the session, one participant was particularly keen to give feedback. He asked about the informal water displays and animated boats, which appear after particularly complex sections, as part of Relations' 'water relaxation' theme. He felt that in contrast to the highly interactive educational aspects, the relaxation screens lacked interaction. In his opinion, an embedded game that allowed users to participate, would be better. An interesting viewpoint, since the water displays had been deliberately designed for users to take a passive break. Perhaps the class of 2006 prefers active relaxation. After the session each participant was given a complimentary *Relations* CD as a token of gratitude and an Easter egg!

4.4 Lessons learned…

There was a clear need to restructure the methodology and task list for the upcoming main study, so as to obtain better information about how participants used their 'non-active' times. Thinking aloud had demonstrated its obstacles and its value (one student did it with aplomb and confidence). We needed to help participants with this and let them have a practice run. There could also be a short demonstration as a guide. Avouris [2001] suggests that think-aloud can be difficult with some users, particularly with young students, for whom verbalization may distract from the tasks. In line with this, we found that the less-experienced participants simply stopped thinking aloud. Research by Boren and Ramey [2000] and Dumas [2003] notes shortcomings in traditional think-aloud methods used in psychological testing by Ericsson and Simon [1993, cited in both the former two sources]. First, there should be a focus not only on the participant's thoughts as they work, but also on their expectations, opinions, interpretations, and anything they wish to report. Second, there is scope for the test administrator to discerningly prompt and probe during task execution, when it is contextually appropriate, but taking care to avoid bias. Should a subject fall silent, the administrator can remind him/her to verbalize. Intervention is also required when there is a system bug, when a subject is 'stuck', or when the subject wants to initiate two-way communication. We also decided to strengthen the informative and ethical aspects of the test plan.

5. THE MAIN STUDY

5.1 Process and participants

A further five volunteers, representing the diverse learner population, were invited to participate in the main study. Table 1 shows their profiles. Meticulous attention was paid to communicating the introductory information upfront. In line with Shneiderman's [1998] requirement and, as in the pilot study, the administrator stressed that the idea was to measure the product, not the person. This was important in order to simulate the real-world situation.

Efforts were made to comply with UNISA's *Human Factors and Ergonomics Society Code of Ethics* [UNISA, 2003] relating to treatment of research participants. They were assured that, if they wished to withdraw, they were not

obliged to continue with the test after the introduction and outline, and that during the process they could terminate the session prematurely at any time. In such cases, they would still receive the promised copy of the *Relations* CD. It is important that participants do not feel they are 'in too deep to quit'. First, such participants could produce incorrect data and second, it would be unethical to put someone in such a situation. After the procedure was explained, all the participants gave informed consent. None of them subsequently took the withdrawal option.

Participant	Gender	Age	Computer experience	Occupation	Qualification	Use computers frequently?
1	Male	37	Advanced	Financial manager	Honours degree	Yes
2	Male	34	Advanced	Computer programmer	Diploma	Yes
3	Male	24	Average	HR consultant	Matric	Yes
4	Female	21	Minimal	Full-time student	Matric	Yes
5	Female	19	Minimal	Full-time student	Matric	Yes

Table 1: Profile of participants in the Main Study

5.2 Methods, tasks and metrics

Reflecting on lessons from the pilot test, changes were made to the test plan and the methods:

- *Fine-grained, self-contained tasks*: The two composite tasks of the pilot test (Section 4.2) were restructured into five finer-grained, self-contained tasks for the main test. This provided more mini-breaks. In addition to the printed task list (Figure 1) and the information on the consent form, clear verbal information was given, explaining what happens in usability testing and what is expected of participants. They were told that they were free to ask questions or to request help during the session.
- *Think-aloud reminders*: The restructuring of the tasks gave the administrator opportunities to remind participants to think out loud. These reminders were not interruptions and were therefore not intrusive.
- *Brief think-aloud training*: Before the actual testing, participants were trained briefly. A 60-second video clip of successful think-aloud usability testing was shown, to help them grasp the expectations of the administrator and to formulate a personal frame of reference. They were also given a think-aloud practice run.
- *Interviews*: After the sessions short interviews were conducted with the participants to gain their overall impressions of *Relations* and their level of satisfaction.
- *Ethical aspects*: These were considered most important, as described in Section 5.1.
- *Standard for comparison*: An issue in usability testing is how to interpret the metrics obtained. What are the norms? Various standards are baselines, benchmarks and best cases. A *baseline* in usability testing, as in other forms of experimentation [Sealander, 2004], is a form of 'pretest', i.e. data established in the absence of any intervention or prior to testing a new system. It serves as a basis of comparison for evaluating effectiveness of a changed scenario. A *benchmark* [Rubin, 1994] is a standard time established either as the average or the maximum time for performing a task. Benchmarks help to evaluate participant performance in a test. *Best case* testing [Rubin, 1994] measures the performance of an experienced user, who is familiar with the type of product. If such a user has trouble, it indicates serious problems in the application. The best case can also be used in a usability specification [Faulkner, 2000] as a target for achievement. In our situation, with a new venture and no standard measures, we took a set of data from an expert user to use as a best-case measure of optimal performance. As the 'expert', we selected the 2006 module leader of *Theoretical Computer Science 1* and asked her to set a realistic standard by working systematically and accurately through Tasks 1, 2 and 3 of the task list shown in Figure 1.

There are five tasks to be completed. Tasks are to be completed in the order given.

Task 1 - From the welcome page go into the main menu.
- In the main menu go to **Properties of relations**.
- You are required to learn about **Reflexivity**.

Task 2 - From the **Properties of relations** menu.
- You are required to learn about **Symmetry**

Task 3 - From the **Properties of relations** menu.
- You are required to learn about **Transitivity**

Task 4 - From the main menu go to **Special kinds of relations**
- You are required to learn about the **Equivalence Relation**

Task 5 From the main menu:
- Go to **Test Yourself** section.

- Read the instruction as to how to answer the questions
- Progress through to the actual Test.
  Here you are only required to fill in the answers related to questions on:
  - Reflexivity
  - Symmetry
  - Transitivity
  - Equivalence

The rest of the answers will be pre-completed.

Figure 1: Task list for usability testing of *Relations*

Tasks 1, 2 and 3 relate to the three major concepts in mathematical relations, reflexivity, symmetry, and transitivity, respectively. Task 4 integrates them in a composite concept. The findings relating to Task 4 follow a similar pattern to the data collected from the other tasks, and are not included in this paper. Task 5, 'Test yourself' is a self-test in *Relations*. Participants were required only to do questions relating to the content they had learned from the assigned tasks; these totaled 20 marks. All the other questions in the test were pre-completed.

Figure 2 is a composite image, showing simultaneous data from three video cameras, as viewed in the observation room of the usability lab. One portion shows the screen as the subject interacts with it. The other two cameras capture the subject's facial expression and mouse-and-keyboard actions, respectively.

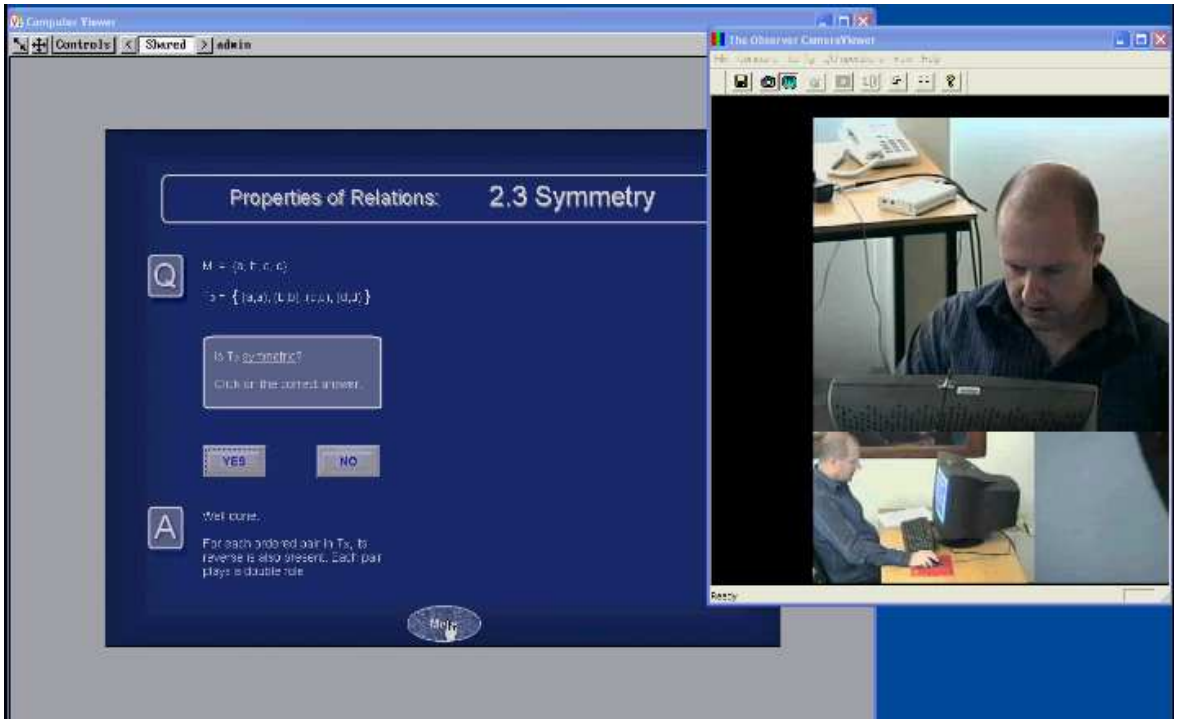

Figure 2: Simultaneous triple-screen video data in the usability lab (used with permission of the participant)

5.3 Findings and discussion

No specific time was allocated to tasks, which contributed to capturing the real user experience. It was notable that some participants repeated certain activities, using *Relations*' user control functionality.

Event logging was not done during the actual testing sessions. It was done afterwards, using the video-recorded data and the think-aloud on audio. Although doing replays takes more time in the laboratory, it allows the test administrator to concentrate unhindered during live sessions. If attention is focused on meticulous live logging and the keying in of events, it might distract the administrator from subtle events in real time, possibly delaying needed intervention. Concurrent observation and accurate logging is a skill that calls for great experience.

Doing *Relations* content and associated exercises							**Test yourself**	
User	Time (secs) navigating by mouse clicks and by using <Enter> key	Time (secs) reading and interacting with the learning content, using <Enter>	Total number of mouse clicks	Cognition time (secs) spent in studying / thinking	Time (secs) typing answers to exercises, including second attempts	Total comple-tion time (secs)	Time (secs)	Score (max 20)
Best case	56	342	60	36	30	534	—	—
1	744	271	56	156	37	1208	119	17
2	154	139	55	0	49	342	25	7
3	250	729	100	341	223	1543	763	13
4	84	449	97	567	172	1272	167	12
5	84 8 (stuck)	533	89	405	89	1119	292	16
Average (excluding best case)	265	424	79.5	294	114	1097	273	13

Table 2: Performance metrics

The performance measurements in Table 2 show the times taken for Tasks 1, 2 and 3, aggregated due to their similarity, and the number of mouse clicks, as well as metrics for Task 5, the self-test. The first line gives the 'best case' performance of the expert, which can be used as an optimal performance benchmark. Mouse-clicking is the way of navigating through menus, supplemented by use of the <Enter> key (times in 2nd column and total clicks in 4th). The reflexivity, symmetry, and transitivity content of *Relations* involve reading theoretical concepts and interacting with or watching animated visuals and worked examples (times in 3rd column). The 5th column is most important, showing cognition time spent on in-depth study/thought. These times were distinguished from other times by analyzing the think-aloud in combination with video. All of this is interspersed with exercises where users type in characters for answers (times in 6th column). Ways of doing the tasks differ from learner to learner in respect of choices to omit or repeat sections, and differ most of all in doing exercises. Users getting an answer wrong received diagnostic feedback. Some repaired their cognitive errors carefully, while others gave a quick incorrect answer, so as to get the solution sooner. The number of clicks is not a straight measure of efficiency, but also indicates aspects such as the amount of repeating, e.g. a learner getting exercises correct and proceeding to the next section would use less clicks, but might re-do learning content sections and examples and click more than average there. No statistical analysis has been done due to the small number of participants [Preece et al, 2003].

The data indicates the interesting presence of an *outlier* [Dumas & Redish, 1999]. This is evidenced by a set of values that is quite different from the others. User **2,** a competent computer programmer, showed a rapid completion time, faster than the *best case*. He was most interested in the procedure, and was primarily aiming to stress-test the system and less keen to do serious learning at that point in time! Note the low 'reading of content' time and zero time spent 'studying'. The test score was also low. *Relations* handled the stress testing well, and the programmer said he would appreciate a further opportunity of using it, just to input garbage and see what happened!

The *best-case* metrics are better than the participants' performances, except in the case of the outlier and some data from User **1.** Since the expert was asked to work systematically and realistically, her best-case data establishes itself as a tentative high-standard benchmark. The averages cannot be viewed as a benchmark due to the low number of cases and the variations in the data, which are addressed next.

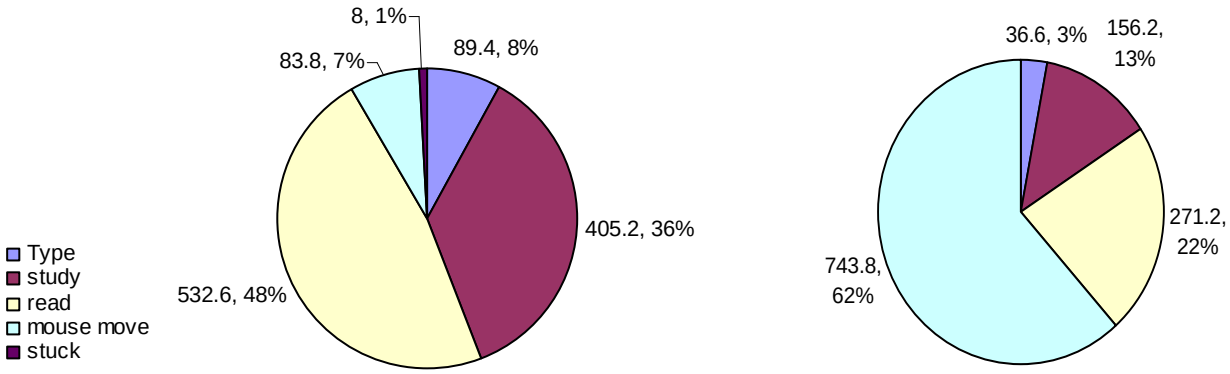

Figure 3: Usage pattern pie-charts

The varying use of time is particularly interesting. Some data from Table 2 is graphically presented in the *usage pattern pie charts* of Figure 3, which show the distribution of time use for two different users, User **5** (left) and User **1** (right). Their 'total times', 1119 and 1208 respectively (see Table 2), are similar and so are their high scores of 16 and 17 for 'Test yourself'. Yet their time usage patterns differ considerably. For example, User **5** moved quickly from one activity/display to another, spending only 7% of the time navigating by the mouse, but 48% of the time reading and 36% studying (i.e. 84% in total on cognitive activities). In contrast, User **1** spent 62% of the time actively navigating, moving backward and forward through content and activities, with the mouse and <Enter>. User **1** spent 22% and 13% of the time respectively on reading and studying (35% on cognitive activities, in contrast with the 84% of User **5**). These usage patterns indicate very different learning styles and activities, yet with similar durations and outcomes.

In one aspect, usage patterns were uniform. The metrics in Table 2 are for Tasks 1, 2 and 3 aggregated. Since the three tasks have homogenous structures, they were combined for analysis. However, for each participant, we also investigated the total time spent on each of the tasks separately. From Task 1 to Task 2 to Task 3 there is a tendency to an increase in complexity and in length, and this was reflected in the time distribution patterns. With one exception the ratio of time spent on the tasks was in the order of 8**:**11**:**15.

User	Successful task completion			Problem issues (no of occurrences)			Number of interruptions by administrator
	Task 1	Task 2	Task 3	Usability errors	Got stuck	Request for help	
1	Yes	Yes	Yes	1	—	—	—
2	Yes	Yes	Yes	1	—	—	2
3	Yes	Yes	Yes	—	—	—	1
4	Yes	Yes	Yes	2	—	—	3
5	Yes	Yes	Yes	—	1	1	1

Table 3: Usability issues: task completion and problems

Table 3 outlines task completion rates, errors and problems identified. *Relations V2* is reasonably mature, having been through different kinds of evaluations, and refined through Versions 2.0 and 2.1 (see Section 2.3). It was appreciated by two cohorts of students in 2004 and 2005 respectively, and is in use again this year. There are very few serious problems and it is easy to use, hence the low scores in the usability error column. All five participants completed Tasks 1, 2 and 3 successfully. With regard to errors, we distinguished between critical and non-critical errors. The latter are errors from which users recover easily and rapidly; they are not shown in Table 3. Regarding more serious errors, three of the five experienced a known weakness in *Relations* [De Villiers 2004], namely that errors occur when students use operations and keystrokes they know from other systems. This finding is in line with Nielsen's [2000] point that five users are sufficient to identify most of the usability problems. The problems encountered were as follows: Users **2** and **4** both clicked on the elaboration block that popped up after a mouse-over, instead of clicking on the menu, while User **1** right-clicked where inappropriate, yet consistently with legitimate right-clicking in other applications. Only one user, **5**, 'got stuck' (see Table 2) and asked for help. The problem was minor and was solved quickly. The 'interruptions' occurred when the administrator stopped users from venturing outside the designated tasks. *Effectiveness* of the application – relating to usability, functionality and screen design – can thus be positively assessed, due to the low error rate, high completion rate, satisfactory test scores and the relaxed progress noted on the videos and think-aloud audios. This is despite the fact that the usability testing addressed some work content which had not, at that stage, been covered in the official academic programme of the module.

To assess user-acceptance and satisfaction, we did not use supplementary questionnaire surveys [Van Greunen & Wesson, 2002] since *Relations* had previously been evaluated this way (Section 2.3). Instead a short semi-structured interview was conducted. Its findings showed high satisfaction levels. All the participants had had enjoyable and satisfying experiences, for reasons such as: '*It (Relations) is interesting*' and '*Ehh.. I just enjoyed it*'. They found the environment very interactive and, as one put it, '*…better than studying from a book, definitely*'. One of the professionals would like a question bank for random question generation, so as to produce different exercises when a learner re-does sections.

6. CONCLUSION AND FUTURE RESEARCH

We first re-visit the research questions posed in Section 3.1 and sum up their answers:

1. In addition to the identification of performance measures and problems, how can testing in a usability laboratory elicit valuable information about cognitive e-learning applications?

We propose that usability testing of e-learning applications should address both the *interfaces* and the *learning content*, because usability and functionality are closely related in e-learning. Performance metrics generated during learning activities can be used towards measuring the *effectiveness* aspect of usability, because success in the cognitive processing induced by learning functionality is fundamental to both usability and utility. It could be argued that conventional usability testing should be conducted on user interfaces only, to investigate users' experiences with navigation features and the menus only. This could lead to improved interaction design, but it would be a paltry use of the capacity of the monitoring and analytical features in usability laboratories.

2. What activities/outputs yield meaningful information about interactive applications in such domains?

Innovative use of the usability lab technology led to *usage analysis* to identify *usage patterns*. Using data from *thinking-aloud by subjects* to clarify how they used their time, a clear distinction emerged between time spent navigating and time spent in cognitive activities. The act of configuring an environment or system to one's own needs is called 'incorporated subversion' by Squires [1999]. In the present context, incorporated subversion of the usability-testing technology led to added-value use of the laboratory and its software in novel and adaptable ways.

3. What notable features emerge from this study that can contribute specifically to a framework for usability testing of interactive e-learning in cognitive domains?

Generic frameworks and methodologies can be established, then customized for optimal usability testing of different kinds of e-learning applications and environments.

- Attention to *ethical* aspects is of vital importance in this close-up recording of personal human activities.
- A *pilot study* is essential to support the sensitive evolving plans and critical judgements that must be made.
- This case study established the value of *thinking out loud* as a source of data, provided it is preceded by adequate *preparation of participants*.
- The more *fine-grained the tasks* selected for testing, the better the data that is recorded.
- Regarding the number of subjects, *five are sufficient to identify usability problems*, but are not enough to conduct serious analysis of learning and cognitive patterns. In this study the five participants generated valuable initial data on usage patterns and learning styles. Future in-depth research should be undertaken on low level, very fine-grained tasks, accompanied by detailed analysis. With more data, realistic averages could be obtained to serve as benchmarks against which to compare future usage studies on small groups or single-subjects.
- For the present, tentative *benchmarks* and *best case* measures were obtained. The best case provides a realistic optimum standard.

This study was not merely a usability evaluation or an examination of the user interface of an e-learning product. It was also a usage analysis of an e-learning application designed to support learning in a cognitive domain. The measurements of cognitive usage provide data that shows the different ways in which different students use its functionality and distribute their time. It is clear that learning style significantly affects the interaction. Over and above capturing performance metrics and identifying problems, novel ways emerged of using the hardware and software capabilities of a usability laboratory to obtain rich information about interactive learning applications.

Nor was it evaluation for the sake of focused usability evaluation of a specific product. The application tested had already been systematically evaluated by various techniques and methods, and has been refined to a state of reasonable maturity. Rather, this case study was an effort to explore usability-testing technology in the e-learning context. We suggest methods, approaches and plans which contribute to a framework for testing usability of interactive e-learning applications in domains where usability − in its broad context − incorporates certain functionality and is closely related to learning content. The standard usability testing conducted for task-oriented products and business systems is not directly relevant to cognitively-oriented applications, such as computer-assisted learning, web-based learning, children's software, and computer games. In the latter cases efficiency − measured by rapid task completion − is not a

meaningful metric, and errors do not necessarily have bad connotations. Just as this study addressed learning applications in cognitive domains, future research can investigate evaluation and testing of other computing- and mobile artifacts in non-standard or unique domains in order to set up appropriate testing frameworks.

We express sincere appreciation to the participants in both case studies − the pilot study and the main study.

7. REFERENCES

ALESSI, S.M. & TROLLIP, S.R. 2001. *Multimedia for learning: methods and development* (3rd ed.). Massachusetts: Pearson Education Company.

AVOURIS, N.M. 2001. An introduction to software usability. Invited paper at Workshop on Software Usability, *Proceedings of 8th Panhellenic Conference on Informatics*, 2: 514-522, Nicosia, November 2001, Livanis Publ., Athens.

BLOOM, B.S. 1956. *Taxonomy of Educational Objectives. Handbook 1: Cognitive Domain.* New York: David McKay.

BOREN, M.T. and RAMEY, J. 2000. Thinking aloud: reconciling theory and practice. *IEEE Transactions on Professional Communication*, 43(3): 261-278.

CEDEFOP (European Centre for the Development of Vocational Training) 2002. E-learning and training in Europe. Luxembourg: Office for Official Publications of the European Communities.

DE VILLIERS, M.R. 2004. Usability evaluation of an e-Learning tutorial: criteria, questions and a case study. In: G. Marsden, P. Kotzé, & A. Adesina-Ojo (Eds), *Fulfilling the promise of ICT. Proceedings of SAICSIT 2004.* ACM International Conference Proceedings Series.

DE VILLIERS, M.R. 2005. e-Learning artefacts: Are they based on learning theory? *Alternation* 12.1b: 345-371.

DE VILLIERS, M.R. 2006. Multi-method evaluations: Case studies of an interactive tutorial and practice system. *Proceedings of InSITE 2006.* In press.

DIX, A., FINLAY, J., ABOWD, G.D. and BEALE, R. 2004. *Human-Computer Interaction*. Pearson Education, Ltd, Harlow.

DUMAS, J.S. 2003. User-based evaluations. In: J.A. Jacko & A. Sears (Eds), *The Human-Computer Interaction Handbook*. Mahwah: Lawrence Erlbaum Associates.

DUMAS, J.S. and REDISH, J.C. 1999. *A practical guide to usability testing*. Exeter: Intellect.

FAULKNER, X. 2000. *Usability Engineering*. Houndsmills: Macmillan Press.

HAYS, P.A. 2004. Case study research. In: K. deMarrais & S.D. Lapan, (Eds), *Foundations for Research*. Mahwah: Lawrence Erlbaum Associates.

ISO 9241. 1997. *Draft International Standard: Ergonomic requirements for office work with visual display terminals (VDT). Part 11: Guidance on Usability*, ISO.

JEFFRIES, R., MILLER, J.R, WHARTON, C. & UYEDA, K.M. 1991. User interface evaluation in the real world: a comparison of four techniques. Proceedings *ACM CHI'91*: 119-124 *Conference*. New Orleans, LA, April 1991.

MAYES, J.T and FOWLER , C.J. Learning technology and usability: a framework for understanding courseware. *Interacting with Computers* 11(5): 485–497.

NIELSEN, J. 2000. Why you only need to test with five users. http://www.useit.com/alertbox/20000319.html. Accessed March 2006.

PREECE, J., ROGERS, Y. and SHARP, H. 2002. *Interaction Design: Beyond Human-Computer Interaction*. John Wiley & Sons, Inc.

PRETORIUS, M.C., CALITZ, A.P. & VAN GREUNEN, D. 2005. The added value of eye tracking in the usability evaluation of a network management tool. In: J. BISHOP & D. KOURIE (Eds), *Research for a Changing World. Proceedings of SAICSIT 2005*. ACM International Conference Proceedings Series.

QUINTANA , C., CARRA, A., KRAJCIK, J. & SOLOWAY, E. (2002). Learner-Centered Design: Reflections and New Directions. In: J.M. Carroll (Ed.), Human-Computer Interaction in the New Millennium. New York: Addison-Wesley.

RUBIN, J. 1994. *Handbook of Usability Testing: How to Plan, Design and Conduct Effective Tests.* New York: John Wiley.

SEALANDER, K.A. 2004. Single-subject experimental research: An overview for practitioners. In: K. deMarrais & S.D. Lapan, (Eds), *Foundations for Research.* Mahwah: Lawrence Erlbaum Associates.

SHNEIDERMAN, 1998. *Designing the User Interface* (3rd ed.). Reading, MA: Addison Wesley Longman.

SQUIRES, D. 1999. Educational software for constructivist learning environments: subversive use and volatile design. *Educational Technology*, 39(3):48-53.

SQUIRES, D. and PREECE, J. 1999. Predicting quality in educational software: Evaluating for learning, usability and the synergy between them. *Interacting with Computers* 11(5): 467–483.

UNISA 2003. Guidelines for ethics in research at UNISA. http://www.unisa.ac.za/cmsys/staff/contents/departments/docs/research_ethical_guidlines.pdf

VAN GREUNEN, D. and WESSON, J L. 2002. Formal usability testing of interactive educational software: A case study. *World Computer Congress (WCC): Stream 9: Usability.* Montreal, Canada, August 2002.

WESSON, J L. 2002. Usability evaluation of web-based learning: An essential ingredient for success, *TelE-Learning 2002*: 357- 363, Montreal, Canada, August 2002.

WESSON, J.L. and COWLEY, N. L. 2003. The challenge of measuring e-learning quality: Some ideas from HCI. *IFIP TC3/WG3.6 Working Conference on Quality Education @ a Distance*: 231-238, Geelong, Australia, February 2003.