How to combine objectives and methods of evaluation in iterative ILE design: lessons learned from designing AMBRE-add

S. Nogry[1], S. Jean-Daubias[2], N. Guin[2*]

[1] PARAGRAPH Lab ,Paris 8 University, Saint Denis, France

[2] LIRIS Lab, Lyon University, CNRS, Lyon, France

Abstract.

This paper deals with evaluating an Interactive Learning Environment during the iterative design process. Various aspects of the system must be assessed and a number of evaluation methods are available. In designing the interactive learning environment AMBRE-add, several techniques were combined to test and refine the system. In particular, we point out the merits of combining measures of the ILE's impact on learning with analyses of the learner's activity. In this paper, the approach used to evaluate AMBRE-add is described and analyzed, and its advantages and disadvantages are discussed. Then a general procedure for evaluating ILEs during iterative design is proposed.

Keywords: *iterative design, usability, utility, acceptance, use patterns, evaluation procedure.*

[*] Sandra Nogry. Email: sandra.nogry@versailles.iufm.fr

Nathalie Guin. Email: *nathalie.Guin@liris.univ-lyon1.fr*

Stéphanie Jean-Daubias. Email: *stephanie.Jean-Daubias@ liris.univ-lyon1.fr*

1. Introduction

For quite some time now, many authors have been pushing for evaluating interactive learning environments (ILEs) early in the development process (e.g. Norman & Draper, 1986; Norman, 1988), in view of catching problems that would be costly if not impossible to correct after the design stage is completed (e.g. Nanard & Nanard, 1998). But how can an ILE be evaluated during its design?

The literature describes the various aspects of these systems that need to be evaluated and proposes a number of techniques for doing so (Kolski, 2001). But as Bastien (2004, p. 461) stressed, the techniques are largely independent of each other, and determining how to combine them is not a simple task. In this paper, we rely on experience gained during evaluations conducted in the AMBRE project to propose a procedure for evaluating ILEs during iterative design. Our procedure is suited to learning environments aimed at enhancing the acquisition of specific skills or concepts by structuring the learner's activity.

Given that the impact of an ILE on learning depends on how the system guides the learner's activity, it seems necessary to analyze both the learner's activity when working on the system and the impact of the system on learning. The procedure we propose is aimed precisely at combining a set of evaluation methods so that the design process can integrate evaluations of the system's usability, assessments of its impact on learning, and analyses of individual learners' activity when working in this type of learning environment.

After presenting the different elements generally evaluated in designing a computer system and some of the methods used to do so, we consider the specific points that need to be addressed in ILE evaluation, and we discuss the applicable methods. Then we present the evaluations we made when designing AMBRE-add, an ILE in which the aim is to learn to solve arithmetic problems; these evaluations, which took place in several stages and involved about a hundred students, were very thorough. The next section presents a critique of our procedure. We state the merits and limitations of the methods used, and then point out the advantages of combining an analysis of the activity carried out in the learning environment with an evaluation of the usability and impact of the system on learning. Lastly, we describe the general procedure we propose here for evaluating ILEs during iterative design.

2. Evaluating an ILE

A set of aspects of the system have to be evaluated. Three of these aspects are particularly important: the usability of the system, its utility, and its acceptance. Usability refers to how easy the system is to use. It is defined as the fit between the way in which a task is carried out by a user and the cognitive abilities of that user (Nielsen, 1993; Farenc, 1997). According to ISO definition 9241-11, a system is "usable" (a) when the user is able to perform the task on the system (effectiveness), (b) when doing so requires a minimal amount of resources (efficiency), and (c) when the system is user-friendly (user satisfaction). Usability is often the only ergonomic criterion considered by designers (Burkhardt & Sperandio, 2004). Senach made the distinction, however, between a system's utility and its usability, where utility is taken to be the fit between the functions offered by a system and those necessary to the user for satisfying the high-level demands of clients (Senach, 1993). As for acceptance, this term refers to users' perceptions of the system's usability and utility, a factor that influences their decision to use or not to use the tool. Acceptance depends on the norms, values, motivations, and affects of users (Nielsen, 1993; Dillon & Morris, 1996; Dillon, 2001).

Specific features of ILEs

ILEs have a number of specific features that must be taken into account in achieving an effective evaluation of their usability, utility, and acceptance. The first concerns the main purpose of any ILE: to promote learning. This objective may be different from the one the learner is striving to reach (a short-term goal). For example, to learn a

particular concept, the user's short-term goal could be to find documents about that concept in a hypermedia. Thus, ILEs can have two objectives at different levels.

The second specificity is related to the ILE's users. Usually, there are two types of users: learners and teachers. Learners are the final users of the system; teachers are its "prescribers" in that they are usually the ones who propose the ILE to their students. Teachers may themselves be secondary users of the system (Dubourg & Teutsch, 1997; Jean-Daubias, 2003) because they may need to set the system parameters to prepare it for use by students. When an ILE is being evaluated, both of these types of users must therefore be considered.

Thirdly, ILEs are often employed in a wide range of situations. They can be used by learners working alone or in pairs, at home or in a computer room with a teacher, or individually in a classroom while other students do other activities. The use of an ILE can be extremely limited (used once) or very frequent in a given period of the year. Given that learning is affected by the situation in which the ILE is used, it is necessary to take this situation into account in assessing any progress made.

Therefore usability-evaluation methods must be adapted to these specific features (Hû & Trigano, 1998, 1999; Jean-Daubias, 2003), and utility-evaluation methods must be geared to the overall purpose of the system, i.e., to promote learning.

Evaluating ILE usability

"The usability of ILEs is contingent upon the quality of the interface (its consistency, readability, how it represents possible actions, etc.), its navigation properties (consistency, simplicity, exhaustivity of possible moves), and its consistency with respect to the goal" (Tricot *et al.*, 2003, p. 396 our translation). Traditionally, two types of usability-evaluation methods are contrasted: analytical evaluation and empirical evaluation.

Analytical evaluation of usability consists of studying the interface relative to a set of standards (e.g. lists of criteria drawn up by ergonomists), in order to determine whether it exhibits certain qualities and to detect any problems it might pose (see for example Bastien & Scapin, 1993; Lewis, Polson, Wharton & Rieman, 1990; Nielsen, 1993; Nielsen & Mack, 1994; Schneiderman, 1992). Although such criteria and attributes can serve as a basis for assessing ILE usability (Squires & Preece, 1999), they have to be adapted to the system's pedagogical goal, its users, the task to be performed, and the theory of learning underlying the tutoring environment (Jean, 2003; Hû & Trigano 1998).

Empirical evaluation consists of collecting data on the behavior of users as they work on the system. The way users utilize the system is observed and analyzed. Detailed individual observations of interactions between a user and the system are conducted to identify each user's skills, detect potential difficulties, or note any unexpected characteristics of the situation (Gagné, Briggs & Wagger, 1988). Empirical evaluation methods are more flexible than analytical ones. They seem to be better suited to assessing ILEs, insofar as the users observed are representative of the target learners.

Evaluating ILE utility

While it is important that an ILE be usable, it is also necessary to make sure that it is useful, i.e., that it serves the purpose for which it was designed. In the case of ILEs, the goal is twofold: learn (the discipline being taught, not how to use the system) and accomplish the tasks proposed by the system in order to achieve that learning (e.g. solve problems, search for information, perform simulations, etc.) (Vivet, 1996; Jean-Daubias, 2003). Although these two levels are connected, there is not always a direct link between accomplishing a task and actual learning; under certain circumstances, failing on a task can even be beneficial to learning. Evaluating the utility of an ILE therefore involves more than just checking to see whether users can carry out the task they wish to accomplish, which can be done using conventional methods; it also involves determining whether the learning goal is reached by the learner.

There are numerous methods for evaluating learning, most of which are empirical. The comparative method developed in cognitive psychology is often chosen for evaluating ILEs (for a description, see Shute & Regian, 1993; for a critical review, see Tricot & Lafontaine, 2002). The problem that arises here is defining the control condition: Should the ILE be compared to oral teaching? Some other system? A shortened version of the ILE being tested? Modular version with components selectively removed for experimental control? Despite this difficulty, the comparative method has been successfully employed to detect system-triggered changes, and to infer - with a certain degree of generality - that knowledge was acquired by learners though interactiob with the ILE. However, such comparisons do not provide insight into what actually happens while the learning device is being used. Online methods (Rouet & Passerault, 1999), on the other hand, supply information about the learner's activity as he/she uses the system. Some online methods are capable of indicating what the learner is focusing on; others concentrate on understanding the learning processes at play. Methods like verbal protocols, interaction records, traces and ethnographic methods can enable one to take the learner activity and the learning situation into account (Barfurth, Basque, Chomienne & Winer, 1994; Fasse & Kolodner, 2000).

Evaluating ILE user acceptance

The third and last dimension to evaluate is the user acceptance of the ILE. Acceptance can be defined as “the demonstrable willingness within a user group to employ information technology for the tasks it is designed to support” (Dillon, 2001). It is the more or less positive attitudes, opinions of the users about the ILE, its usability and its utility (Tricot *et al.*, 2003). Whether individual or collective, this mental representation is thought to condition the decision of whether or not to use the ILE, i.e., the decision of the prescribing teachers to propose or not to propose it, and the motivation of learners to use it. As stated by these authors, "Acceptance can be contingent upon factors like the culture and values of users, their affects, their motivation, and the social organization and practices into which the ILE fits to a greater or lesser degree" (Tricot *et al.*, 2003, p. 194, our translation).

According to Nielsen's model (Nielsen, 1993), the acceptance of a system has both a practical dimension and a social dimension. Usability and utility are among the criteria that determine the system's practical acceptance; they are necessary conditions for proper acceptance, although in most studies usability and utility are evaluated before acceptance.

The present review is not exhaustive. The problem of evaluating ILEs is addressed more comprehensively in several other recent papers (Delozanne, 2006; Tricot *et al.*, 2003).

3. What methods of evaluation to use during the iterative design of AMBRE-add

As stressed in the preceding section, it is particularly important in evaluating an ILE to verify whether it is usable, useful and acceptable by both the prescribing teachers and the learners. Software engineers have developed design methods that include evaluation in the design cycle. The iterative design method consists of gradually perfecting system specifications by repeatedly evaluating the system and then making the required modifications until a satisfactory product is obtained (see also Delozanne, 2006). Usability-evaluation methods are aimed specifically at providing information to the system designer, so they are meant for the design cycle. By contrast, methods for assessing an ILE's impact on learning are often developed for *a posteriori* evaluations of ILE performance, i.e., once the design process is complete. Yet, to improve a system being designed, evaluation methods must not only provide information about the system's performance, but should also point out its shortcomings and their causes (De Vries, 2001).

This section is aimed at showing how to take into account these different facets at an early stage of the design process of AMBRE-add. It first presents the basic principle underlying AMBRE-add and its domain of application. Then the various evaluations carried out during design and the merits and limitations of each one are presented.

The AMBRE-add learning environment

AMBRE-add, an interactive learning environment developed as part of the AMBRE project (Guin-Duclosson Jean-Daubias & Nogry, 2002), is designed to guide learners in solving arithmetic. The problems AMBRE-add proposes to learners are one-step arithmetic problems, which are studied by 7- and 8-year-olds in the second and third grades of French elementary school. The problems describe a concrete situation such as a playing with marbles. For example, "Pierre has 32 marbles. After recess, he has 45 marbles. How many marbles did he win during recess?"

Certain arithmetic problems can be correctly solved by preschoolers, whereas others still cause trouble at the end of ninth grade (Marthe, 1982); the difficulties elementary school children have in solving arithmetic problems stem essentially from their inability to correctly represent the situation described in the problem statement (Greeno & Riley, 1987). Many studies have shown that the problem category and the type of unknown are two factors that enter into determining problem difficulty.

The AMBRE project postulates that the fact of asking learners to reformulate the new problem and then look for a similar problem already seen in order to adapt its solution to the new problem, can teach them to model the problem and acquire the corresponding problem class (Vergnaud, 1982; Riley, Greeno & Heller, 1983; Guin, 1991) and its solving techniques. So, the system shows the learners work-out examples (typical problems), and then helps them solve a new problem by taking them through the following steps of the Case-Based Reasoning cycle (CBR; see Figure 1):

- After reading the problem statement, the learner reformulates the problem in order to identify the elements needed to solve it.
- Then, he/she is given a set of typical problems already seen and has to choose a problem similar to the one to be solved.
- Next, he/she has to adapt the solution of the chosen problem to the current problem.
- Finally, he/she classifies the new problem, associating it with one of typical problem already seen.

Figure 1. The AMBRE cycle

Thus, learners carry out the steps themselves but are guided by the system. On each step, a diagnosis is made by the system, which evaluates the learner's work. Explanations are proposed whenever necessary to help the learner understand the mistakes made and find the correct answer.

Design and evaluation cycle of AMBRE-add

AMBRE-add was developed by a multidisciplinary research team based on an iterative design cycle (Figure 2). Partners with varied profiles (researchers in computer science, in cognitive psychology and in the didactics of mathematics, a pedagogical advisor, teachers, and learners) participated in the AMBRE-add design, following a differentiated design method (Jean-Daubias, 2004, 2009).

Figure 2. The iterative design cycle of AMBRE-add showing the gradual expansion of users, tests, and system validations (taken from a diagram in Jean, 2000).

First, initial general specifications defining the principle of the AMBRE project were proposed by the design team (see Figure 2). These specifications were then applied to the domain of arithmetic problems in order to design and implement the AMBRE-add learning environment. The system was evaluated from the technical standpoint by the developers. Next, evaluations by inspection were carried out, some by the designers, with the help of the pedagogical advisor and teachers. Each evaluation was followed by new specifications and modification of the system. Once the design process produced a version of the ILE that met the project's initial objectives and could (*a priori*) be incorporated into existing teaching methods, four empirical evaluations were carried out. Each evaluation was followed by new specifications and modification of the system.

According to Nielsen (1993), usability and utility are necessary conditions for proper acceptance. So we assumed that the system's usability and utility had to be evaluated first. A first laboratory evaluation was performed with five second graders. Then a comparative evaluation with three second-grade classes was carried out in school during six sessions, in order to evaluate the impact of its basic principle (the AMBRE cycle) on learning (Nogry, 2005). According to the results of this second evaluation, which were not satisfying, two other evaluations were carried out to test the system on third graders. For each evaluation conducted, we first present objectives, the techniques employed and a summary of the results.

Evaluation by inspection

First, the design team conducted analytical evaluations of several aspects of the system. In addition to a technical evaluation, we assessed system usability by inspecting different facets of the interface while paying particular attention to interface consistency, guidance, and error management (Bastien & Scapin, 1993; Jean-Daubias, 2003; Schneiderman, 1992). This ergonomic evaluation of usability was completed by a pedagogical inspection conducted by teachers and the pedagogical advisor of our team .They analyzed whether the activities proposed by the system in each step were suited to second and third graders. This inspection was also aimed at verifying the instructions, the error messages and the arithmetical problems in order to verify how well they matched the learners' level. However, it did not enable us to eliminate all usability problems.

Empirical evaluation conducted in a laboratory

The first empirical evaluation, conducted in a laboratory, was aimed at discovering the greatest difficulties experienced by second-grade children when using the AMBRE-add learning environment. As recommended by Nielsen and Landauer (Nielsen & Landauer, 1993), five children were observed; they worked individually on the ILE for one session lasting 45 minutes. After the session, the children were asked various questions about the use of the system. The observation was based on ergonomic criteria selected among those proposed by Bastien and Scapin (1993), Nielsen (1993), and Schneiderman (1992), and included system learnability, efficiency, error management, cognitive overload, and satisfaction. For each criterion, a set of observables were defined in advance.

This initial observation and analysis of the children's answers to the questions allowed us to describe their use of the system, and to identify the usability problems they encountered. For example, some difficulties understanding the

general principle of the system were noted. As well, an arithmetic-related difficulty was exhibited by all users in the adaptation step (see Jean-Daubias, 2004 for a description). Note that this difficulty was not mentioned during the pedagogical inspection.

However, the short duration of the observation did not separate difficulties related to learning how to use the system from problems that persisted after the first session. Besides, the children often asked the observer for assistance rather than using the help provided by the system (even though she tried to limit her mediations). This undoubtedly reduced the reliability of our observations about the help and diagnostic capabilities of the system.

In the light of these observations, we drew up a list of recommendations for modifying certain parts of the interface, for revising the system's display format in order to improve its learnability and understandability, and for changing the problem-adaptation step in a way that would eliminate the arithmetic difficulty uncovered here.

Evaluation of the system's impact on learning

This first evaluation was centered on usability assessment. We supplemented it with a second evaluation using the comparative method in order to test the impact of the AMBRE cycle (the five step proposed by the system to guide the learner) on the learning of the problem classes and their solving techniques. AMBRE-add was compared to two control systems (termed the reformulation and solving system and the simple solving system) in order to show that guided problem solving via the AMBRE cycle steps (see Figure 4) has a greater impact on learning than other forms of computer-assisted arithmetic problem solving. Given that this impact is linked directly to how the system structures the learner's activity during use, the comparative approach was combined with activity-description techniques.

Evaluation design

Three 2nd Grade classes participated to the study. Each pupil (N=76; means age: 7 years five months) had to use one of the three systems during 6 sessions. The first control system, the "simple-solving system" (see Figure 3), displayed worked-out examples first and then went directly on to solving a new problem without guidance. In this condition, the solving process had far fewer steps than in AMBRE-add. So an additional task was proposed which consisted of searching information in a problem statement.

The “reformulation and solving” system presented solved problems and guided the learner in solving the new problem. The learners reformulated the problem and then wrote the solution. Then they read the problem report. In contrast to Ambre-add, this system does not allow the child to select and use a typical problem. We chose this control condition in order to verify the impact on learning of choosing a typical problem and adapting it. The two “control” interfaces were very similar to the Ambre-add interface.

Figure 3. The three systems used in the experiment

This experimental approach was supplemented by recordings of interaction traces (number of solved problems, time spent on each step, types of errors made), observations, a questionnaire, and a discussion. The observations enabled us to describe how the system was actually used and any interactions between learners and observers (like questions asked by learners, difficulties encountered, the type of help given). After the last session, a questionnaire and a group discussion were proposed in order to find out what the learners thought about the systems (e.g. difficulties experienced and general satisfaction).

Main results

The analysis of the problem-solving test scores (Table 1) indicated a significant improvement in performance after working on a system, no matter which one ($F(3,192) = 18.1$, $p < 0.001$). The type of system did not have a significant effect ($F(2,64) < 1$, ns). The system and test variables did not interact ($F(6,192) = 1.15$, $p = 0.33$), which means that pupils who worked on AMBRE-add did not make more progress than pupils who worked on a control system.

Table 1. Mean number of problems correctly solved on each test, as a function of the computer system used

	Pre-test		Test after the 4th session		Post-test		Delayed post-test	
	M	SD	M	SD	M	SD	M	SD
AMBRE-add	2.38	1.94	4.04	1.95	3.81	2.14	4.16	2.04
Simple solving system	2.76	1.61	3.75	2.13	4.08	2.04	3.59	1.79
Reformulation & solving system	2.42	1.38	3.33	2.06	3.75	1.96	4.42	1.16

According to these findings, the AMBRE cycle had the same impact as the other solving methods (for a more detailed presentation and discussion of the results, see Nogry, 2005).

The interaction traces recorded by the different systems showed that the children solved fewer problems with AMBRE-add than with the simple-solving system (an average of nine vs. fourteen problems solved over the six sessions). Our observations showed that children had difficulties in reading the instructions, and in understanding the help and explanation messages. AMBRE-add users often had difficulties in finding the source of their errors. Some of them adopted a passive attitude toward the system (they didn't try to correct their mistakes) or looked for the solution by trial and error. These results pointed out a number of system usability and acceptance problems for second graders: some aspects of AMBRE-add do not seem to be well suited to children in this grade, who are unskilled readers and not always able to work on their own.

Merits and limitations of the evaluation methods used

Note first of all that this evaluation proved to be very time-consuming, not only in the preparation phase (design and development of the control systems), but also in the testing and analysis phases.

The goal of the **comparative approach** was to make sure that any progress made was indeed the result of the AMBRE principle rather than of carrying out other activities. However, the activities carried out on the three systems differed, from the standpoint of the problem solving method employed, of the number of problem solved and of the experimenter interactions that took place (the questions asked by the children as they worked were not the same for the three systems). This reduced the comparability of the situations and thereby limited the validity of the quantitative results. This situation poses the problem of the control condition. Trying to test the effect of the AMBRE cycle on learning before testing the impact of AMBRE-add on learning itself seemed to be too ambitious.

In this stage of the design, the **techniques used to characterize the learners' activity** during system use (observation, noting questions asked, interaction traces) were the most informative methods for the design process. Our analyses of the observations made and the questions asked not only told us how much time was needed to learn to use the ILE, but also pointed out persistent difficulties and allowed us to describe certain aspects of actual system use. Regarding **observation of learners' activity**, as seen above, the information gathered as the learners worked on the different systems, and the questions they asked, turned out to be critical in understanding the results of the quantitative analyses. However, contrary to what we had planned, it wasn't always possible to observe the pupils' activity for extended periods; during the first few sessions, for instance, the experimenters were called upon frequently so they had little

time to write down their observations. The **interaction traces** recorded by the system therefore acted as a good supplement to the observations. These data are easy to obtain and are indicative of the generality and reliability of observations made by experimenters.

Recommendations

Based on this evaluation, we modified AMBRE-add in a number of ways: the help and diagnostic messages were made easier to understand and an animated pedagogical agent with a synthesized voice was added to the system, both to limit reading and to play a motivational role (Lester et al., 1997). In addition, some simple activities were included for familiarizing the children with the environment or to propose remedies geared to their particular difficulties. Furthermore, we reconsidered the targeted audience for AMBRE-add: this ILE seems to be more suitable for third graders.

Evaluation with third graders

After the above experiment which pointed out AMBRE-add's lack of acceptance for second graders, we carried out the next evaluation on third graders. In France, the third-grade curriculum still includes one-step arithmetic problems like the ones in AMBRE-add, because pupils at this level continue to have trouble "modelling" some of them. However, third graders are better readers than second graders, and work better on their own. The purpose of the present experiment was to assess the usability and acceptance of AMBRE-add for third graders, in view of comparing the way the learners actually used the system to the way it was designed to be used, and also to detect difficulties experienced by this user audience.

Method

In this experiment, 21 third graders (N=23; mean age: 8 years 8 months) worked on AMBRE-add individually in the computer room of their school. The arithmetic problems presented by the system were adapted to the pupils' Grade: the numerical values and the type of the unknown were modified, the easiest problem classes were not presented. During the four 45-minute sessions we observed their actual use of the system; during each session, the experimenters took turns observing the children. Over the four sessions, every child was observed at least once. After the fourth session, the pupils were questioned individually about what strategy they had used during the adaptation step (the key step of the AMBRE cycle). Interaction traces were also compiled by the system. They also had to fill out a questionnaire about any difficulties experienced, their comprehension of the vocabulary words used in the interface, and their degree of satisfaction.

Results

As expected, the third-grade pupils worked well alone and had no trouble reading the instructions or the help and diagnostic messages. They also had no problems operating the system. By the second session, they barely asked any more questions: they had learned how to use the system in a single session. The analysis of the interaction traces showed that they solved the problems faster than the second graders had.

The third graders also seem to have resorted less often to trial and error. However, navigation difficulties and problems understanding messages on the reformulation step persisted for certain pupils. Based on our analysis of these difficulties, we designed a tutorial better suited to these learners. The new tutorial provided more information about how to navigate in the system and made the ILE's general principle easier to understand. This evaluation also

encouraged us to test the system's learning effect on third graders, for AMBRE-add seems to be well-suited to this user audience.

Merits and limitations of this evaluation

In the experiments with third graders, more detailed analyses of the learners' activity were conducted with more systematic observations of the learners. The number of methods employed was reduced by concentrating on individual observation, an analysis of interaction traces, and a questionnaire. This evaluation setup had the advantage of being easy to implement with a small number of classes, while still affording valuable information for the design process. These analyses enabled us to compare the system's actual use with its prescribed use, and to identify certain difficulties as well. It enabled us to take stock of the behaviors characteristically adopted by learners in the course of each session and to see how those behaviors evolved. Note that these methods pose some methodological problems: To what level of granularity should learner activity be analyzed? What interactive actions are relevant and meaningful for learners? Besides, it said nothing about the progress made by learners working on the system.

Comparative evaluation with third graders

The encouraging observations made with third-grade pupils, along with the teachers' interest in the solving procedure and the problems proposed by the system, encouraged us to test the learning impact of AMBRE-add in this user audience. The purpose of this study is to evaluate if the system itself actually supported learning and that the progress made wasn't due to the mere fact of repeating the tests. We did so by conducting a comparative study combined with more systematic observation methods and an analysis of the children's interaction traces. In the control condition pupils simply carried out another activity in class with the teacher.

Method

Participants were twenty-three third graders (mean age: 8 years 2 months). In this evaluation, the children took an initial test (T1) consisting of solving six arithmetic problems (Figure 4). Then half of the children in each class worked on AMBRE-add in the computer room four 45-minute sessions, while the other half stayed in the classroom with the teacher to carry out another activity unrelated to problem solving. After the four sessions, the pupils took a second test (T2) made up of six problems similar to the ones on the first test. Next, the two groups were switched, that is, those children who had not yet worked on the system did so for four sessions, while the others stayed in class (Figure 4). After these four sessions, another post-test was run (T3). This experimental design was chosen in order to control the differences between subjects.

Several learners were observed individually during the sessions. A grid for quantifying various predefined behaviors (proposed by Baker *et al.*, 2004, based on Lloyd & Loper, 1986) was filled in based on three observations per session of each pupil. Finally, a questionnaire was answered by the pupils after the last computer session.

Figure 4. Experimental design of experiment in a third-grade class

Results

First of all, the learners' activity as they worked on the system confirmed the observations made in the preceding evaluation. For example, the behavior analysis showed that only a small number of pupils adopted a trial-and-error approach. The two conditions were compared (AMBRE-add vs. control) combining the results obtained within each condition and then an analysis of variance was conducted on the correct answers, with *computer system* (AMBRE-add

vs. control condition) (table 2) and *test* (before vs. after) as within-subject variables[†]. This analysis yielded a *test* effect ($F(1,20) = 10.4$, $p < .005$) and a marginally significant interaction between *system* and *test* ($F(1,20) = 3.7$, $p = .070$): More progress was made in the AMBRE-add condition than in the control condition. AMBRE-add thus seems to promote learning among third-grade children. However, an analysis of the pupils' individual results showed that the best learners (error rate lower than 20% at pre-test) and pupils signalled by teacher to have a lot of trouble made no progress at all.

Table 2. Mean number of problems correctly solved, by condition and test

	Test 1		Test 2		Test 3	
	M	SD	M	SD	M	SD
Group 1	3.00	0.86	4.27	2.08	4.64	1.93
Group 2	4.00	2.05	4.20	1.81	5.10	1.20

To help such pupils, the activities proposed by an ILE should be adapted to their level. The present evaluation confirmed the importance of incorporating a model of the learner into the system. In addition, we planned to design additional remedial work activities in order to eliminate specific problems.

Merits and limitations of this evaluation

The results of the two experiments conducted in third grade classes were satisfying from the designers' point of view. They suggest that the AMBRE-add learning environment is usable (easy to learn and understand), useful (enables pupils to solve arithmetic problems and is a source of progress), and acceptable to third-grade pupils.

From the methodological standpoint, the individual observations made and the quantitative observation grid used for this evaluation enabled us to characterize the learners' activity and the difficulties encountered as they worked in this learning environment. In addition, we were able to measure the system's impact on learning via the comparative method we set up. Note, however, that the findings would have been more robust if the experiment had been run on a larger number of classes.

The control condition chosen here was easier to implement yet was sufficient for testing the system's impact on learning. We were nonetheless fully aware that an evaluation of this type could not tell us whether the learning was the result of the principle proposed by AMBRE-add -- guiding the learner through the solving steps of the AMBRE cycle -- or the result of repeated practice at solving problems. However, we felt it was necessary to verify the impact of the system itself before assessing the effect of its underlying principle by comparing it to systems proposing other solving techniques.

Conclusion

In summary, in the course of the iterative design cycle of the learning environment AMBRE-add, we performed four evaluations with second or third graders: the first in a laboratory and the next three in school. The evaluations pointed out problems of various types, on the basis of which we revised the system, added some functionalities, and reconsidered the user audience to target. In our last experiment on third graders, the combination of a comparative approach and an activity analysis allowed us to relate the learners' progress to the activities they performed and the difficulties they came across during the sessions. This combination of methods seems to be promising for gaining insight into the reasons why progress is or isn't made, and for developing new design specifications accordingly. It can

[†] The fact of counterbalancing the order for working on the systems can be considered as an experimental precaution rather than a variable. Note that in this analysis, the scores obtained on the second test were considered twice.

aid in identifying the causes of difficulties experienced by ILE users, and can thereby help the design team to find alternative solutions. The merits of analyzing instrument-mediated activity and use patterns in the course of the design process has already been demonstrated in cognitive ergonomics (see Rabardel, 2001; Decortis *et al.*, 2001; Trouche, 2002). In the domain of ILEs, however, only marketed systems or already available websites have been analyzed at the present time. Analyzing instrument-mediated activity for the purposes of improving ILE design is a topic that was not approached until recently (see Brassac, Gregori & Hautecouverture., 2007; Cottier & Choquet, 2005) and is still not widespread. These considerations are taken up in the next section, where we present our proposal of an ILE evaluation procedure.

4. What evaluation procedure should be used in iterative design?

What conclusions can be drawn from the evaluations carried out during the iterative design of AMBRE-add? To what extent can the findings of these studies be generalized?

As we have seen, different techniques can be implemented during iterative ILE design to evaluate its usability, its utility in learning the targeted knowledge, and its acceptance. No consensus has been reached on how to combine these methods. For this reason, we devised a procedure for evaluating AMBRE-add during iterative design.

As Littman and Soloway stressed, "There can be no doubt that evaluating ITS's is costly, frustrating and time-consuming" (Littman & Soloway, 1988). However, not all evaluation techniques have the same cost. Implementing our evaluation procedure allowed us to specify the merits and limitations of the various techniques used. On this basis, we developed an evaluation procedure for use in the iterative design of an ILE for individual learning.

In this section we begin by making some recommendations drawn from our evaluation experience and backed by theoretical considerations, in order to provide some guidelines for choosing a good evaluation procedure that will enhance ILE design. These recommendations are followed by a proposal for a procedure we think is well-suited to assessing school-oriented interactive learning environments during the iterative design process.

Recall first of all that designing technical devices can be seen as an initially broad and poorly delineated problem that is approached by gradually narrowing down the alternatives: the design team starts with a large number of degrees of freedom, which decreases little by little until the choices made become increasingly irreversible (De Terssac, 1996). The design of ILEs falls right in line with this structure, so that choosing an order for conducting evaluations and deciding how to integrate the results of the evaluations into the system's design have a substantial impact on the final product.

Three dimensions of a system must be evaluated: usability, utility, and acceptance. According to Nielsen's classic model, a system's usability is a prerequisite for its utility, which in turn is a prerequisite for its acceptance (Nielsen, 1993). In this model, an evaluation procedure must assess usability first, utility second, and acceptance last. However, assuming that utility is necessary for system acceptance does not imply that this property alone is enough to ensure acceptance (Dillon & Morris, 1996). What should be done if, at an advanced stage in the design process, the evaluation of a usable and useful system indicates low acceptance? To avoid such a predicament, Tricot *et al.* (2003) suggested that acceptance should be evaluated at the same time as the other dimensions. Our evaluation experiments on the AMBRE-add learning environment led to the same conclusion. Although we had not initially planned to formally evaluate acceptance in our experiment with second graders, we realized then that it was important to take this dimension into account: our observations of system use in a school setting provided us with critical issues about the system's acceptance.

The cost of implementing an evaluation is another criterion that can be used as a basis for choosing what methods to employ. However, cost must be considered relative to the benefits, for the design process, of the information obtained from the evaluation. For example, even if a classroom evaluation involving several sessions is costly, the usability and acceptance of an ILE designed for use in the schools cannot be appraised unless such an evaluation is made.

On the basis of these considerations, the procedure we propose combines various techniques for assessing a system's usability, utility, and acceptance and incorporating the findings into the iterative design process as early as possible. This procedure is intended for use in designing interactive learning environments that structure the learner's activity in order to promote the acquisition of new skills or knowledge, especially ones designed for individual use in a school setting.

After defining the specifications of the ILE and making an initial prototype, the next step is to perform an evaluation by inspection. This method has the advantage of being easy to implement early in the design process, even when not all of the system's functionalities have been programmed. It can also be useful in reorienting the design process at little cost. The evaluation can deal with the system's usability, based on the many pre-existing criteria; with its pedagogical or didactic aspects, which will require the help of experts in these domains; or with the system's acceptance *a priori*. Based on the outcome of the evaluation, new specifications will need to be drawn up. This loop is continued until a technically "stable" system is obtained.

Empirical evaluations conducted in a laboratory are necessary at this point. Observing the way representative users work on the system, accompanied by individual interviews, can uncover major weaknesses in usability and perhaps also point out difficulties of a didactic nature from which new specifications are drawn up and the prototype modified. Repeated use of the system over several sessions can prove useful for going beyond the initial system-learning phase and finding out how learners work on the system over a longer period. This evaluation can be reiterated several times if the first evaluation leads to substantial system modifications.

The next step consists of observing and analyzing the activities of learners as they work on the system in a school setting. Clearly, the use patterns of a technical device like an ILE do not depend solely on the tool itself, but also result from the learners' schemes of use (Rabardel, 1995; 2001); these in turn interact with the existing practices and characteristics of the educational institution where the system is in use (Hussenot, 2005; Folcher, 2003). An activity analysis can provide information about a system's usability (e.g. initial ILE learning time or persisting difficulties), its acceptance (e.g. changes in motivation or incompatibilities with the user's time schedule or computer-room location), and/or its use patterns and how they evolve from one session to the next. It can also point to individual differences in system use and appropriation. The analysis of instrument-mediated activity, and more broadly of use patterns, can rely on observations, analysis of interaction traces, and interviews that take into account the subject's point of view about his/her own activity (motivation, meaning given to actions). This analysis can lead to modifications in the system and the design of a teaching scenario to make the system easier to learn, to appropriate and to provide guidelines for using it (Hautecouverture, Gregori & Brassac, 2007; Trouche, 2002). It can be repeated in the course of the design cycle in order to fine-tune the product based on regular observations of its use patterns (Hautecouverture, Gregori & Brassac, 2007).

The activity analysis should be combined as soon as possible with an analysis of the knowledge or skill acquired by learners working on the system. Relating the enhancement of individual knowledge and skills to the activities carried out in the learning environment can provide insight into exactly what actions are associated with learning. This analysis can look at changes in the way subjects respond when interacting with the ILE, or at scores on domain-specific tests taken before, during, and after system use.

The choice of what type of activity analysis to conduct, and what types of tests to run to measure learning, must be specific to the ILE's domain of application or to the knowledge at stake, and these choices must be in keeping with how far along the design process has come. In the initial phases, an inductive analysis (not based on any assumptions) conducted via an observation/analysis device with few constraints, can bring out important, unexpected phenomena and point to potential modifications that need to be made in the system. Àt this point, tests aimed at measuring learning can be used as a supplement for discovering possible tendencies rather than for obtaining exact measures and reliable statistical differences.

To increase the validity of the methods used for assessing knowledge or skill acquisition, a control condition should be set up. Choosing an effective control condition can be problematic, and a few guidelines are proposed here for making this choice. To begin, comparing the ILE with a no-intervention situation (see Ainsworth *et al.*, 1998) can offer useful information about the system's utility. If the tests proposed are appropriate, and if progress is observed, then the comparison can serve to ensure that the effect obtained is not due to some other activity carried out (e.g. with the teacher). A comparative analysis can also be set up for other purposes, such as testing the impact of a particular system function. Here, the ILE is compared to another version of the same system (Aleven. Koedinger & Cross, 1999; Aleven & Koedinger, 2002; Luckin *et al.*, 2001), in which case it isn't so much the utility of the system itself that is at stake, but the utility of one of its functionalities. However, as mentioned above for our own experiment, defining another version of an ILE can be a touchy undertaking, and to be able to interpret the results of the comparison, it is important to look carefully at the activity the other version demands.

Finally, it should be noted that in many cases, a comparative approach is used to see how well a system under design measures up to an existing system or to conventional teaching by a teacher (see Shute & Glaser, 1990; Koedinger *et al.*, 1997; Meyer, Miller, Steuck & Kretschmer, 1999). We think that such comparisons are more suitable for evaluations conducted near the end or after the design process rather than during design. This type of evaluation is aimed at demonstrating the superiority of the ILE over existing systems or methods for teaching similar material (e.g., it triggers faster learning, better learning, or the acquisition of additional knowledge or skills).

To conclude, let us mention that any evaluation procedure obviously has to be made to meet the design team's objectives and requirements. Our proposed procedure seems to be applicable to various domains, provided the methods used to evaluate learning - which are entirely dependent upon the knowledge or skills to be acquired on the ILE - are adapted to the particular learning domain.

Conclusion

Given the multiple aspects of an interactive learning environment that need to be assessed, and the many methods available for doing so, what evaluation procedure should be implemented during the design of such a system?

We were faced with this question when designing the AMBRE-add learning environment. Our approach consisted of evaluating the usability and the utility of the system by drawing upon existing research in order to conduct a number of evaluations combining various methods. In this article, we described the available methods and then gave a critical review of our evaluations during AMBRE-add design. We demonstrated the merits for design -- and also the limitations -- of the various methods used. This critical analysis allowed us to point out the advantages of examining learner activity during the system use at an early stage of the design process, and to propose an evaluation procedure based on the findings.

The procedure begins with an evaluation of the ILE's usability using the methods described in the field of human-machine interaction. Then the activity of learners is analyzed over several sessions in a real-world situation, in order to complete the usability evaluation, identify the different patterns of system use, and collect data on the system's

acceptance. The use-pattern analysis can be combined with an evaluation of the ILE's impact on learning, in view of determining what patterns of system use are associated with learning. Based on each evaluation, new specifications are written up for developing a revised version of the system, via an iterative design process.

Acknowledgments

This research was supported by the multidisciplinary program STIC-SHS "Société de l'Information" of the French National Research Center (CNRS). We would like to thank the computer scientists who developed the AMBRE-add system and the teachers who participated in the study. We are also grateful to the three reviewers for their constructive remarks.

Notes on contributors

Sandra Nogry is a researcher affiliated with the Comprehension, Reasoning, and Acquisition of Knowledge Team at the Laboratory "Paragraphe" (University Paris 8). She also works with the Imaging and Information Systems Laboratory LIRIS (UMR 5205. Her research deals with the evaluation of ILEs, design, learning, and analogical problem solving.

Address: Université Paris 8, 2 rue de la liberté, 93526 Saint-Denis Cedex, France

E-mail: sandra.nogry@versailles.iufm.fr

Web: http://paragraphe.crac.free.fr/articles.php?lng=fr&pg=111

Stéphanie Jean-Daubias is a computer science professor at Claude Bernard University in Lyon. She is affiliated with the Imaging and Information Systems Laboratory LIRIS (UMR 5205) as a member of the SILEX team (Supporting Interaction and Learning by Experience). Her research is conducted as part of the PERLEA and AMBRE projects and deals with the personalization of learning in view of providing teachers with tools for individualized learning.

Address: Université de Lyon, CNRS - Université Lyon 1, LIRIS, UMR5205, F-69622, France

E-mail: Stephanie.Jean-Daubias@liris.univ-lyon1.fr

Web: http://liris.cnrs.fr/stephanie.jean-daubias/

Nathalie Guin is a computer science professor at Lyon 1 University. She is affiliated with the Imaging and Information Systems Laboratory LIRIS (UMR 5205) as a member of the SILEX team (Supporting Interaction and Learning by Experience). Her research deals with knowledge-based systems for ILEs, and her current main topics of interest are related to ILE personalization: interpretation of activity traces, learner profiles, and user-geared activities. All of these research themes are studied within the AMBRE project.

Address: Université de Lyon, CNRS - Université Lyon 1, LIRIS, UMR5205, F-69622, France

E-mail: Nathalie.Guin@liris.univ-lyon1.fr

Web: http://liris.cnrs.fr/nathalie.guin

References

Ainsworth, S. E., Wood, D. & O'Malley, C. (1998). There is more than one way to solve a problem: Evaluating a learning environment that supports the development of children's multiplication skills, *Learning and Instruction*, 8 (2), 141-157.

Aleven, V. & Koedinger K. R. (2002). An Effective Meta-cognitive Strategy: Learning by Doing and Explaining with a Computer-Based Cognitive Tutor. *Cognitive Science*, 26 (2), 147-179.

Aleven, V., Koedinger, K. R. & Cross K. (1999). Tutoring answer explanation fosters learning with understanding. *Proceedings of AIED-99*, (pp. 199-206). Amsterdam, Netherland: IOS Press

Baker, R.S., Corbett, A.T &, Koedinger, K.R. (2004). Detecting Student Misuse of Intelligent Tutoring Systems. *Proceedings of ITS'2004* (pp.531-540). Maceio, Brasil: Springer.

Bastien, C. & Scapin, D. (2004). La conception de logiciels interactifs centrés sur l'utilisateur: étapes et méthodes. *In* Falzon P. (Ed.), *Ergonomie* (451-462). Paris: PUF.

Bastien, C. & Scapin, D. (1993). Ergonomic criteria for the evaluation of Human-Computer Interface, *Technical report No. 156*, INRIA.

Barfurth, M.A., Basque, J., Chomienne, M. & Winer, L.R. (1994). Les instruments de collecte de données de recherche qualitative dans des environnements pédagogiques informatisés. *In* Bordeleau P. (Ed.), *Apprendre dans des environnements pédagogiques informatisés* (485-548) Editions Logiques.

Burkhardt, J.-M. & Sperandio, J.-C. (2004). Ergonomie et conception informatique. *In* Falzon P. (Ed.). *Ergonomie*, (437-450) Paris: PUF.

Bruillard, E., & Baron, G.-L. (2006). Usages en milieu scolaire: caractérisation, observation et évaluation. *In* Grandbastien M., Labat J.-M. (Eds), *Environnements informatiques pour l'apprentissage humain,* (Chapter 12). Paris : Hermès-Lavoisier.

Cottier, P. & Choquet, C. (2005). De l'usager construit à l'usager participant. *Proceeding of EIAH'2005* (pp.449-454). Montpellier, France: Hermes.

Decortis, F., Daele, L., Polazzi, L., Rizzo, A. & Saudelli B. (2001). Nouveaux instruments actifs et activités narratives. *Revue d'Interaction Homme-Machine*, 2 (2), 1-30.

Delozanne, E. (2006), Interfaces en EIAH, *In* Grandbastien M., Labat J.-M. (Eds.), *Environnements informatiques pour l'apprentissage humain*, (pp. 223-244). Paris: Hermès-Lavoisier,

De terssac, G. (1996). Le travail de conception: de quoi parle-t-on? *In* De Terssac G., Friedberg E. (Eds), *Coopération et conception* (1-22). Paris : Octares.

de Vries, E. (2001). Les logiciels d'apprentissage, panoplie ou éventail? *Revue Française de Pédagogie*, 137, 105-116.

Dillon, A. & Morris M. (1996) .User acceptance of information technology: theories and models. *In* Williams M. (Ed.), *Annual Reviewof Information Science and Technology*, Vol. 31. Medford, NJ: Information Today.

Dillon, A. (2001) User Acceptance of Information Technology. In W. Karwowski (Ed). *Encyclopedia of Human Factors and Ergonomic*s. London: Taylor and Francis.

Dubourg, X. & Teutsch, P. (1997). Interface Design Issues in Interactive Learning Environments. *Proceedings of IFIP WG 3.3 Working Conference: Human-Computer Interaction and Educational Tools*. Sozopol, Bulgaria.

Farenc, N. (1997). ERGOVAL: une méthode de structuration des règles ergonomiques permettant l'évaluation automatique d'interfaces graphiques. PHD dissertation, University Toulouse I.

Fasse, B.B. & Kolodner, J.L. (2000). Evaluating Classroom Practices Using Qualitative Research Methods: Defining and Refining the Process. *Proceedings* of the International *Conference of the Learning Sciences* (pp. 93-198).

Folcher V. (2003). Appropriating artifacts as instruments: when design-for-use meets design-in-use. *Interacting With Computers, 15*, 647-663.

Gagné, R.M., Briggs, L.J. & Wager, W.W. (1988). *Principles of instructional design*. New York: Holt, Reinhart, & Winston.

Greeno, J.G. & Riley, M.S. (1987). Processes and development of understanding. *In* Weinert F.E., Kluwe R.H. (Eds), *Metacognition, motivation and understanding*, (289-313). Lawrence Erlbaum Associates.

Guin, D. (1991). La notion d'opérateur dans une modélisation cognitive de la compréhension des problèmes additifs. *Math. Inf. Sci. Hum*. 113, 5-33.

Guin-Duclosson, N., Jean-Daubias, S. & Nogry, S. (2002). The Ambre ILE: How to Use Case-Based Reasoning to Teach Methods. In *Proceedings of ITS'2002*, pp. 782-791. Biarritz, France: Springer.

Hautecouverture, J.-C., Grégori, N. & Brassac, C. (2007). Appropriation d'une plate-forme de Coopération par des enfants en cadre scolaire. *European Review of Applied Psychology*, 57 (1), 1-16.

Hoecker, D. & Elias, G. (1986). User evaluation of the LISP intelligent tutoring system. *In Proceedings of the human factors society*, *32 (3)*, 313-324.

Hû, O., & Trigano, P. (1999). Considerating Subjectivity in Software Evaluation - Application for Teachware Evaluation. In Vanderdonckt, J. & Puerta, A. (Ed.*), CADUI 99, **C**omputer-Aided **D**esign of **U**ser **I**nterfaces* (pp. 331-336), Kluwer Academic Publisher, 1999.

Hû, O., & Trigano, P. (1998). Propositions de critères d'évaluation de l'interface homme-machine des logiciels multimédias pédagogiques. *Proceedings of IHM'98*. Nantes.

Hussenot, A. (2005). Trajectoires d'usage d'une solution TIC: traduction, "enaction" et appropriation. Proceedings of the *3rd GDR TIC et Société*, Paris.

Jean, S. (2000). PÉPITE : un système d'assistance au diagnostic de compétences, Thèse de doctorat de l'Université du Maine.

Jean-Daubias, S. (2003). Vers une définition des spécificités des EIAH dédiés à l'évaluation pour l'application de recommandations ergonomiques. *Revue d'Interaction Homme-Machine*, *4 (1),* 535-538.

Jean-Daubias S. (2004). De l'intégration de chercheurs, d'experts, d'enseignants et d'apprenants à la conception d'EIAH, *Proceedings of TICE 2004* (pp. 290-297). Compiègne.

Jean-Daubias S. (2009). Differentiated design: a design method for ILE. Research *report LIRIS2009-015*.
http://liris.cnrs.fr/Documents/Liris-4131.pdf

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, *8*, 30-43.

Kolski, C. (2001), *Analyse et conception de l'IHM: interaction homme-machine par les systèmes d'information.* Paris : Hermès.

Lester, J.C., Converse, S.A., Kahler, S.E., Barlow, S.T., Stone, B.A. & Bhoga, R.S. (1997). The persona effect: affective impact of animated agents. In *proceedings of Computer-Human Interaction'1997*. http://www.sigchi.org/chi97/proceedings/paper/jl.htm

Lewis, C., Polson, P.G., Wharton, C. & Rieman, J. (1990). Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. In *Proceedings of CHI'90: Human Factors in Computing Systems* (pp. 235-242). New York: ACM.

Litmann, D. & Soloway, E., (1988). Evaluating ITSs: The cognitive science perspective. *In* Polson M. & Richardson J. J. (Eds), *Foundations of Intelligent Tutoring Systems*. Hillsdale, NJ: LEA.

Lloyd, J.W. & Loper, A.B. (1986). Measurement and Evaluation of Task-Related Learning Behavior: Attention to Task and Metacognition. *School Psychology Review*, *15 (3)*, 336-345.

Luckin, R., Plowman, L., Laurillard, D., Stratfold, M. & Taylor, J. (2001). Narrative evolution: learning from students' talk about species variation. *International Journal of AIED*, *12*, 100-123.

Marthe, P. (1982). Problèmes de type additif et appropriation par l'élève des groupes additifs Z et D entiers relatifs et décimaux relatifs. *Thèse de doctorat, EHESS*, Paris.

Meyer, T. N., Miller, T. M., Steuck, K. & Kretschmer M. (1999). A multi-year large-scale field study of a learner controlled intelligent tutoring system. In Lajoie S., Vivet M. (Eds), *Proceedings of AIED'99* (pp. 191-198). Amsterdam: IOS Press.

Nanard, J. & Nanard, M. (1998). La conception d'hypermédias, *In* Tricot A., Rouet J.-F. (Eds), *Les hypermédias, approches cognitives et ergonomiques,* pp. 15-34. Paris : Hermès.

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

Nielsen, J. & Landauer ,T.K. (1993). A mathematical model of the finding of usability problems. In *Proceedings of ACM INTERCHI'93 Conference* (pp. 206-213). Amsterdam: The Netherlands.

Nielsen, J. & Mack, R. L. (Eds.) (1994). *Usability inspection methods***.** New York, NY: John Wiley & Sons.

Nogry, S. (2005). Faciliter l'apprentissage à partir d'exemples en situation de résolution de problèmes - Application au projet AMBRE. *Thèse de doctorat de l'Université Lumière Lyon 2*.

Nogry, S., Jean-Daubias, S. & Duclosson N. (2004). ITS evaluation in classroom: the case of the AMBRE-AWP. In *Proceedings of ITS 2004* (pp. 511-520). Maceio, Brasil : Springer.

Norman, D. A. & Draper, S. (1986). *User-Centered System Design*. Erlbaum: Hillsdale NJ.

Norman, D. A. (1988). *The Psychology of Everyday Things*. Basic Books: New York.

Rabardel, P. (1995). *Les hommes et les technologies, approche cognitive des instruments contemporains*. Paris: Colin.

Rabardel, P., (2001) - Instrument mediated activity In Situations. In Ann Blandford, Jean Vanderdonckt and Phil Gray (Eds.), *People and Computers XV -Interactions Without Frontiers*, pp. 17-30. Springer-Verlag.

Riley, M.S., Greeno, J.G. & Heller, J.I. (1983). Development of children's problem-solving ability in arithmetic. In , Ginsburg H.P. (Ed.), *The development of mathematical thinking*. New York: Academic Press.

Rouet, J.-F., & Passerault, J.-M. (1999). Analyzing learner hypermedia interaction: An overview of online methods. *Instructional Science*, *27*, 201-219.

Sander, E. & Richard, J.-F. (1997). Analogical transfer as guided by an abstraction process: the case of learning by doing text editing. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 23, 1459-1483.

Schneiderman, B. (1992). *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Reading, MA: Addison-Wesley.

Senach, B. (1993). L'évaluation ergonomique des interfaces homme — machine. *In* Sperandio J.-C. (Ed.), *L'ergonomie dans la conception des projets informatiques,* pp. 69-122. Paris: Octares.

Shute, V.J. & Regian J. W. (1993). Principles for Evaluating Intelligent Tutoring Systems. *Journal of Artificial Intelligence and Education*, *4 (2/3),* 245-271.

Shute, V. J. & Glaser, R. (1990). A large-scale evaluation of an intelligent discovery world: Smithtown. *Interactive Learning Environments, 1*, 51-77.

Squires, D., & Preece, J. (1999). Predicting quality in educational software: Evaluating for learning, usability, and the synergy between them. *Interacting with Computer*, *11(5)*, 467-483.

Tchounikine, P. (2002). Pour une ingénierie des environnement informatiques pour l'apprentissage humain. *Revue Information-Interaction-Intelligence, 2(1),* 59-95.

Tricot, A., Plégat-Soutjis, F., Camps, J.-F., Amiel, A., Lutz, G. & Morcillo, A. (2003). Utilité, utilisabilité, acceptabilité: interpréter les relations entre trois dimensions de l'évaluation des EIAH. In Desmoulins C., Marquet P., Bouhineau D. (Eds), *Proceedings of EIAH 2003* (pp. 391-402). Strasbourg, France: Hermes.

Tricot, A. & Lafontaine, J. (2002). Une méthode pour évaluer conjointement l'utilisation un outil multimédia et l'apprentissage réalisé avec celui-ci. *Le Français dans le Monde*, 41-52.

Trouche, L. (2002). Genèses instrumentales, aspects individuels et collectifs. *In* , Guin N., Trouche L. (Eds), *Calculatrices symboliques - Transformer un outil en un instrument du travail mathématique: un problème didactique* pp. 243-275. La pensée sauvage édition.

Vergnaud, G. (1982). A classification of cognitive tasks and operations of the thought involved in addition and substraction problems. *In* Carpenter P.T., Moser, J.M., Romberg, T.A. (Eds), *Addition and substraction: A cognitive perspective*, pp.39-58. Hillsdale: Erlbaum.

Vivet, M. (1996). Evaluating Educational Technologies: Evaluation of Teaching Material Versus Evaluation of Learning? In *Proceedings of CALISCE* (pp. 37-38). San Sebastian, Spain.