IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Hershey PA 17033-1117, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITP4107

Evaluating Multi-Media Educational Software: The Dast Experience

Fawzi Albalooshi and Eshaa M. Alkhalifa
Department of Computer Science, College of Science, University of Bahrain, P. O. Box 32038, Isa Town, Bahrain
Tel: (973) 449999, Fax: (973) 682582, fawzii@batelco.com.bh, ealkhalifa@sci.uob.bh

ABSTRACT

In recent years computer based instruction has become increasingly popular as a teaching tool. Research in this area questions the reliability of such tools and their ability to transfer knowledge. This introduces a need for proper evaluation methodologies of Multi-Media education software, to be constructed to ensure that they are indeed achieving their purpose. This paper looks at the traditionally followed approaches in the evaluation of this type of software. A demonstration is made of how evaluation can be achieved with the participation of all those involved with using the system. The paper also demonstrates the applicability of this technique to systems by showing a case study of a Data Structures Tutorial Package (DAST).

INTRODUCTION

Computer-based instruction provides educators with a powerful technological tool to aid them in reaching their teaching objectives. Recent advancements in multimedia made it possible to incorporate sound and animation into the same presentation, clearly provide more means for information transfer than classroom whiteboards and textbooks. This tool may even aid in re-enforcing student learning as well as overcoming traditional problems that commonly exist with the traditional approaches. However, many researchers including Beatte, K. (1994), Reiser, R. and Kegelmann, H. (1994), believe that educational software must be evaluated to ensure its teaching benefits on the learners before being approved for use. Questions such as "Do the students like the software" and "Who's using it?" are inadequate as a measure of effectiveness. What is being emphasized is the most fundamental evaluative question: "What's being learned by the students?" A good evaluation must establish whether this type of representation is able to overcome a particular learning problem, and then follows that by a deeper search to investigate the nature of the learning experience and its benefits to students.

The evaluation procedure proposed here emphasizes the participation of all parties involved in the evaluation process, such as, educators, technical experts, and the target learners. Background information is first collected about the package content and its technical performance in addition to finding a method through which the effect of the CBI on student's learning outcome is measured. The collected evaluation information then is analyzed in rigorous detail to determine the suitability of the CBI under analysis as a teaching medium. The presented procedure is applied to evaluate a Data Structure Tutorial Package (DAST) used to teach undergraduate students the concepts related to the stack abstract data type.

TRADITIONAL EVALUATION PROCEDURE

A review carried out by Reiser et al. (1994) showed that in most cases the people who take part in the evaluation were teachers that had to go through the software similar to a student, and then fill out a rating form by comparing the system to what would occur in a classical classroom session. Usually a wide variety of the CBI features are reviewed including, content, technical characteristics, documentation, instructional design, learning considerations, software objectives, and the handling of social issues. Only a small number of evaluators gathered evidence to demonstrate the effectiveness of the CBI in teaching. The authors concluded that organizations should incorporate students as participants in the evaluation process, and that they should be assessed on how much they learned as a result of using the software. Beattie (1994) also suggested a number of evaluation techniques, some of which are pilot testing, before/after testing, expert criticism, and student questionnaires. Tam, M., Wedd, S., and Mckerchar, M. (1997) went one step further, when they proposed a three parts evaluation procedure including peer review, student evaluation, pre- and post-testing. On the other hand, some evaluators concentrated on the effectiveness of the use of particular media as opposed to another. Pane J., Corbett A., and John B. (1996), for example, examined the impact of computer-based animations and simulations on student's understanding in time-varying biological processes. They setup two student groups based on prior test performance in the course to compare computer based and paper based instruction, using as main measure for comparison the pre- and post- test results. Further tests of the animation presentations was attempted by Byrne M., Catrambone R., and Stasko J. (1999) who examined whether animations would help students learn computer algorithms more effectively. Their approach was mainly based on pre and post testing the student groups participating in the experiments. If the last experiment had accents on effectiveness, the one for Lawrence A., Badre A., and Stasko J. (1994) concentrated on finding the difference in student performance in carefully selected pre and post tests including differentiating between declarative and procedural questions.

Although the importance of evaluation as a vital player in any instructional software is evident to all researchers, there do not exist any guidelines through which such evaluations could take place. An example of a problem that may exist, is the series of experiments that were aimed at testing the differences in instructional effectiveness of the animation versus textual media. These tests depended on providing a clear sequence of photographs to show the procedure while in the animated versions, the animation was shown on a screen. Jennifer Freyd (1987) showed through a large number of experiments the basis of what she called "representational momentum". This theory explains a natural tendency to treat any series of images, as equivalent to an animation and vice versa.

Therefore, comparing the two media through tests of effectiveness may not result in any desirable results because what is estimated does not indicate the difference in "cognitive load" during the learning process. Students learning from these textbooks may learn as effectively as the ones that learn through animation, but end up with a smaller overall efficiency when their learning rate is measured by time. The techniques proposed here aim at setting some ground rules through which the evaluation of Multi-Media systems do not fall into these pitfalls.

CASE STUDY: EVALUATING DAST

The Data Structure Tutorial system (DAST) is a Multi-Media tutoring system that was developed at the department of Computer Science in the University of Bahrain. The system aims at teaching and or re-

This paper appears in *Issues and Trends of Information Technology Management in Contemporary Organizations,* the proceedings of the Information Resources Management Association International Conference.

enforcing the basic concepts of the Data Structures course by presenting its content in a combined animated, and textual mediums simultaneously. The aim of the evaluation procedure was to test the suitability of this CBI in meeting its targets and for this the evaluation and analysis was done in three main stages. A pre-evaluation stage was to place participating students in three groups and prepare the pre- and post- tests and the evaluation survey. Implementing the evaluation procedure and data collection followed. The final stage was to analyze the data and determine the effect of the tutorial system on the students' learning.

A Pre-Evaluation Step

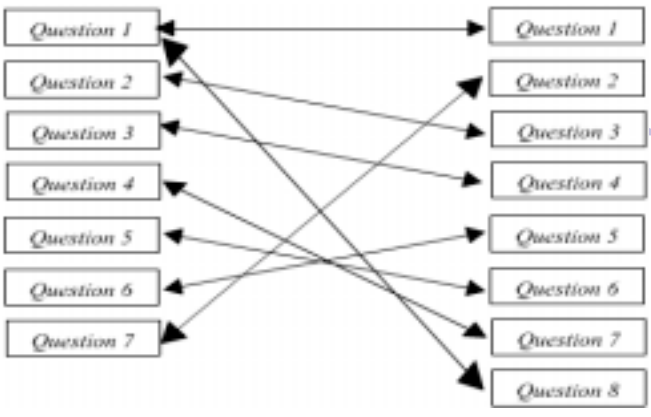
The evaluation procedure must go in stages starting with the development of the system itself and ending with the tests of how effective it is in transferring the information to students. This system started out with a thorough analysis of the educational content, which was done by peer review of the content and examples included in the system. Students were pre-tested to assess their levels in comparison to each other with respect to their learning abilities and then separated into three groups of 15 students by keeping the means of their test grades equal. The aim of this step is to allow an even distribution of students in the groups based on their learning abilities. Without this test, it is quite possible that by accident, a group exposed to a particular condition may have students who are sharper and more capable of learning than another group. Therefore, this step was essential to isolate the variable of student ability from the comparison table.

Two tests were then prepared composed of seven and eight questions to be exact where one of the questions in the first test was broken up into two in the second. The questions were carefully written to ensure that each question on the first test mapped exactly onto one or more on the second test to allow for comparisons on a question-to-question basis to check for differences in students levels within particular domains. Byrne et al. (1999) for example found that the use of interact ional animation improved student responses to procedural questions, while Lawrence et al. (1994) made similar findings simply through interactive laboratory sessions.

An example of the mapping implemented in this particular case can be seen in these two questions: "List and explain the data variables that are associated with the stack and needed to operate on it?" and "List the data variables and operations associated with the stack?" The first would appear as question number 4 on the first day, and the second would appear as question number 3 on the second day. A sample diagram of the mapping is as shown in figure 1.

In addition to these tests an evaluation questionnaire was prepared to allow students to highlight any weak or strong areas they found while interacting with it. In a sense, this would allow students to take an active role in the evaluation process and describe their point of view. Caffarella (1987) proposed some guidelines for such an evaluation form and these have been adopted in the form presented here. Most questions require subjective judgment to the effectiveness of the

Figure 1: A sample mapping of the questions between the post-test and the pre-test



CBI program and how capable it is in meeting its education goals; including questions about program goals, content, audience, instructional strategies, design, appropriateness, etc. The aim of this step is to complement the objective results of testing student performance on the pre and post evaluation tests.

Implementation of the Evaluation Procedure

The evaluation procedure concentrated on testing all possible conditions, which implied that students had to be tested if they used the system only, attended the classroom session only or attended the classroom followed by the system. The first two conditions would test for the differences in the effectiveness of teaching in a classroom versus through a Multi-Media system. The third condition would then be compared to the two above by showing if the system provides any re-enforcement to the learning level attained after a regular classroom and if so how much re-enforcement resulted. In general the technique followed is shown in Figure 2.

Figure 2: The evaluation procedure

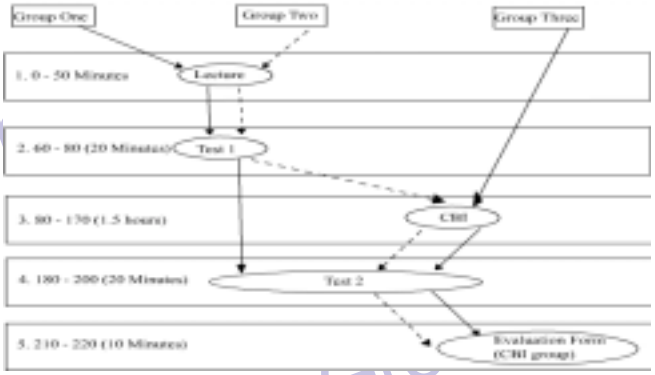


Table 1: Total student scores in pre and post tests

Group One		Group Two		Group Three	
Test One	Test Two	Test One	Test Two	Test One	Test Two
6	4.5	9	16.5		12.5
8	12.5	9	11		8.5
8.5	8	11.5	14		10.5
7	8	8	10.5		6.5
12	13	7.5	7.5		13
22	24	6	14		13.5
10	19	12	13		9.5
9	8.5	8.5	14		12.5
8.5	6	5	10		8.5
17	18.5	11.5	18		6
6	8	7	7.5		12.5
8.5	9.5	5	9.5		13.5
11.5	13.5	8	12		11.5
8.5	6.5	16.5	15		10.5
7.5	7.5	10	12.5		16.5
SD= 2.923	SD: 2.880	SD= 4.167	SD= 5.305		SD= 2.759
X>10.5= 4	X >10.5=6	X>10.5= 4	X>10.5=11		X>10.5=10
X<10.5=11	X <10.5=9	X<10.5=11	X <10.5=4		X <10.5=5

The three groups followed the paths shown. Note that the variation between the two tests allows them to be directly mapped onto each other without any serious problems. They were, however, reworded to reduce the chances of students remembering the answers of the first test. Therefore, the only factor of difference was the time duration between the two types of presentation of material and the use of the system versus the classroom lecture. Students were allowed to take their time during a lab session that usually takes approximately

two hours of which they took at most the period shown of 1.5 hrs. Additionally, steps 1 and 2 of the evaluation procedure took place on day one and steps 3,4 and 5 took place on the following day. To allow students to rest following the classroom lecture and to allow them to forget the questions they were asked in the first test. Students were not informed that they would be retested on the following day, so they were unaware of the events of the following day and had no reason to wish to retain any information concerning the first test. Evaluation questions were asked to experts who have previously taught the course, as well as to students who were subject to this experiment.

Analysis of the Results

In order to avoid falling into the pitfalls of previous systems a thorough analysis of the results was carried out with respect to all the data that was collected. The student marks in the two tests that were given were analyzed using the Analysis of Variance (ANOVA) test. This test allow one to test the difference between the means by placing all the data into one number, which is F, and returning as a result one p for the null hypothesis. It will also compare the variability that is observed between conditions, to the variability observed within each condition. The static F is obtained as a ratio of two estimates of the students' variances. If the ratio is sufficiently larger than 1, the observed differences among the obtained means are described as being statistically significant. The term of null hypothesis represents an investigation done between samples of the groups and with the aim that learning was not a product of the treatment. In order to conduct a significance test, it is necessary to know the sampling distribution of F given that the significance level needed to investigate the null hypothesis. It must also be mentioned that the range of variation of averages is given by the standard deviation of the estimated means. Student scores and the Standard Deviation (SD) according to their groups were as shown in table 1.

The aim of this data is to arrive at some measure of the effectiveness of this system without falling into the temptation of comparing the effects of the different media with each other. Therefore, the benchmark used here and is proposed in general is to compare this type of CBI with a regular classroom lecture. The ANOVA tests did indeed show that there is a significant improvement in Group Two between the first test that was taken after the lecture and the second test that was taken after using the system. However, this was not sufficient to be able to pinpoint the strengths of the system.

Therefore, a more detailed analysis was done of the student performance in the individual questions of test one and test two. Since the questions were mapped onto each other by design as is shown in Figure 1, it was easy to identify significant changes in student grades in a particular question types in the students of group two who responded to similar questions before and after the use of the system. An example is a highly significant improvement with $F=58$ and $p<.000$ was observed in the question "Using an example, explain the stack concept and its possible use?" Which is an indication that the use of the system did strongly impact the student understanding of the concept of a "stack" in a functional manner. Another advantage for mapping the questions as was done is the ability to compare between subjects in groups one and three who either attended the lecture only or used the system only. Oddly enough, the same question comes up again only this time with favor towards the classroom only case and with an $F=5.02$ and $p<.03$. This raises a serious question that would not be raised if not for the evaluation design assumed here. The question is: if the system did worse than classroom lecture in that particular question, then how can they so strongly re-enforce the understanding of the concept following the classroom lecture.

Another point of view is to examine the scores by using the total average, which is 10.639, therefore approximating the border-line becomes 10.5 and the rest of the scores will be divided around this line. It is to be noticed that the scores of the third group were not so high, but most of them were over the average and comparing with the second group, shows the results are close even if the third group took only the CBI package while group two had both lecture and CBI package learning. It also underlines how much the second group improved their test results after taking the CBI and in the same time showing that the first group had not improved much only with the lecture learning. Alkhalifa (2001) showed through two experiments that the difference in error levels between students who are exposed to a logical task may significantly differ whether the task is a "moving" system or a stationary one. Without the use of this particular statistic, the advantages of having animation in this multi-media system would not fully be assessed.

These results seem to indicate that the use of the system may introduce a "limiting" effect that follows the initial introduction to the concepts Albalooshi and Alkhalifa (2001). Classroom lectures introduce students to the concepts allowing them all the freedom to select all types of applications, which is in some ways overwhelming. The use of the system, on the other hand, produces a safe haven to test their ideas and strongly pursue the examples they can imagine. It goes without saying that such a conclusion would have been impossible to reach if the questions were not purposely set in the shown mapped fashion.

Incorporating Student Opinion

Students of groups two and three who were exposed to the system, were asked to fill in an evaluation form composed of a series of questions as proposed by Caffarella (1987). The questions were broken up into several main sections with two to three questions in each. These include;

1. Program goals:- This includes questions to see if students understood the aim of the CBI.
2. Program content: - This includes one question requesting students to judge if the CBI is in line with the University taught materials.
3. Audience for the CBI Program: - This section includes four questions about the suitability of the CBI to this particular group.
4. Instructional Strategies: - This section includes two questions about the suitability of the CBI's approach to teaching and if it can be stopped at any time.
5. Program Design: - This section has six question about feedback, speed of presentation, user control, the use of graphics, sound, etc and readability issues.
6. Appropriate Use of Computers: - Two questions here ask students if this application is appropriate to be presented on a computer and if it takes advantage of the interactivity offered by computers.
7. Program Techniques: - This section has four questions that ask about issues related to software execution including the clarity of directions and programming errors if any.
8. Cost/Benefit Analysis: - Two questions ask about the required time a student needs to use this system and if it is worth the investment.
9. Overall Evaluation: - Questions concerning listing the software's strengths; weaknesses; the user's overall evaluation on a scale; and whether they believe the University should adopt it.

With respect to the DAST system, students in general gave ratings, of around 4 to 5 on a scale that went 0 to 6 with the highest for "The use of graphics, sound, and color contributes to the student's achievement of the objectives" and "The user can control the sequence of topics within the CBI program." The lowest score was 3.559 for "The level of difficulty is appropriate for you". Therefore, it seems that the students in general enjoyed learning through the system although they found the level of difficulty of the concepts presented challenging.

In addition, to all this, three peer experts filled in evaluations forms to rate the system from an instructor's point of view and they gave the system an average rating of 5.33 on the same scale of 0 to 6.

CONCLUSION

It goes without saying that software evaluation is critical to any educator and when the concerns about the effectiveness of the evaluation procedures grow as cognitive researches pinpoints inadequacies in presumptions one has to turn to an "overall view" of the system. Although the DAST system presented has both an animation module, as well as a textual module within the same framework, the individual

testing of each seemed fruitless. Past studies found the two to be equivalent Lawrence et al. (1994) and Byrne et al. (1999) because their methods of evaluation concentrated on comparing the two media with respect to performance. Freyd (1987) showed through several experiments that a comparison of this type is fruitless because the human mind seemed to cognitively interpret one form into another. However, this does not inform us of anything concerning the cognitive load on memory and whether it enhances learning to use one or the other. This implies that separating the media in testing is not informative at all when evaluating a CBI as a whole because results can be misleading. On the other hand, we have shown here that when all the types of different media are presumed to be one united system, and evaluating it from an instructor's, a student's, as well as a performance perspective results are much clearer. The framework used for this evaluation was clearly presented showing why the mapping of pre- and post- test questions is essential to the success of this procedure. Another point worth mentioning is the equal distribution of questions between procedural and declarative with the addition of a conceptual question that combines both. These tools allowed a deeper analysis of the system that gave interesting insights into the points of the strength of the CBI. It is these little surprises that everyone seeks when evaluating the abilities of CBI programs to help guide future designs.

REFERENCES

Alkhalifa, E. M. (2001). Directional Thought Effect in the Selection Task. Proceedings of the Third International Conference on Cognitive Science. Beijing, China, pp. 171-176, August, 2001.

Albalooshi, F., & Alkhalifa, E. M. (2001). Multi-Media as a Cognitive Tool: Towards a Multi-Media ITS. Proceedings of IEEE International Conference on Advanced Learning Technologies: Issues, Achievements and Challenges. Madison, Wisconsin. pp 231-234 August 2001.

Beatte, K. (1994). How to Avoid Inadequate Evaluation of Software for Learning. IFIP Transaction A [Computer Science and Technology]. Volume A-59, pages 245-258. Melbourne, Australia.

Byrne, M. D., Catrambone, R., & Stasko, J. T.(1999). Evaluating animation as student aids in learning computer algorithms, Computers and Education, v. 33(4) pp. 253-278.

Caffarella, E. P. (1987). Evaluating the New Generation of Computer-Based Instructional Software. Educational Technology, 27(4), 19-24.

Department of Computer Science, McMaster University, Hamilton, Ontario, Canada (1999). Abstract Data Type Demonstration Web Site [online]. Available: http://www.cas.mcmaster.ca/cs3ea3/1997/ (October 12th, 1999).

Freyd, J. (1987). Dynamic Mental Representations. Psychological Review, v.94(4), pp.427-438.

Lawrence, W., Badre, A. N., & Stasko, J. T. (1994). Empirically Evaluating the Use of Animation to Teach Algorithms (Technical Report GIT-GVU-94-07), Georgia Institute of Technology, Atlanta, 1994.

Pane, J. F., Corbett, A. T., & John B. E. (1996), Assessing Dynamics in Computer-Based Instruction, In Proceedings of the 1996 ACM SIGCHI Conference on Human Factors in Computing Systems. Vacouver, B.C. Canada, April 1996, pp.197-204.

Reiser, R.A., & Kegelmann H. W. (1994). Evaluating Instructional Software: A Review and Critique of Current Methods. Educational Technology, Research and Development. Volume 42, Part 3, pages 63-69, USA.

Shaffer, C. A., Heath, L. S., & Jun Yang (1996). Using the Swan Data Structure Visualization System for Computer Science Education. SIGCSE Bulletin, Volume 28, Part1, pages 140-144.

Shaffer, C. (1997). Swan [online]. Available: http://geosim.cs.vt.edu/Swan/Swan.html (October 12th, 1999).

Tam, M., Wedd, S., & McKerchar, M. (1997). Development and Evaluation of a Computer-Based Learning Pilot Project for Teaching of Holistic Accounting Concepts. Australian Journal of Educational Technology. Volume 13, part 1, pages 54-67, Australia.