Evaluating The Educational Impact of An Intelligent Algebra Tutor Software: A Research Proposal

Haibei Zhang

Abstract

Intelligent Tutoring System (ITS) shows its effectiveness in improving performance of conceptual and procedural knowledge for students ranging from elementary school to college level. Empirical researches show that ITS surpasses traditional Computer Aided Instruction (CAI) and can be a competitive alternative of human-led tutoring. Experimental research and survey research are two ways to carry out an ITS evaluation research. Based on the context of Ms. Lindquist, a new algebra tutor, I design in this proposal a research to evaluate its effectiveness in education and several factors that may affect it.

Introduction

Intelligent tutoring systems are increasingly being employed in education. (Major & Wood, 1996) Intelligent Tutoring Systems (ITS) are computer programs that are designed to incorporate techniques from the Artificial Intelligence (AI) community in order to provide tutors which know what they teach, who they teach and how to teach it. AI attempts to produce a behavior in computer, which if performed by a human would be described as "good teaching" (Elsom-Cook, 1987). A current ITS typically has four components: expert knowledge module, student knowledge module, tutoring module, and user interface. (Nwana, 1990) On the opposite of constructivist theory which stresses "learning by doing" or "learning by design"(Duffy & Jonassen, 1992), ITS designers stress "learning by being told." (Michalski & Chilausky, 1980) Due to this nature of ITS, it shows certain advantages over traditional labs or discovery learning (Albacete, 2000) such as: clear articulation of knowledge, providing clear diagnosis of errors of students, showing how and why certain instructional techniques work or not. (Nwana, 1990) These features enable ITS to be a perfect test-bed for many theories and an ideal alternative of human tutors offering one-to-one tutoring, which was proven to be a very efficient way to attain higher achievements. In one research, 98% of students with private tutors performed better than those without private tutors. (Bloom, 1984).

Ms. Lindquist is a dialog based ITS. It is developed by Neil Heffernan and Ken Koedinger (2001) at Carnegie Mellon University. It initially asks the student an algebra word question, in which there are several relations between several variables. If the student's answer is correct, the ITS will jump to the next question. If the answer is wrong,

the ITS will restate the question in which some variables and relations are omitted, such that the question is shorter and simpler. The ITS also provides appropriate hints upon student's wrong answer. It is intelligent and looks like a human expert because the tutoring it provides is based on student's response.

Instructional software, like all other educational material, should be evaluated before it is used in the classroom or research laboratory. (Heller,1991) The purpose of this research is to investigate the effectiveness of a specific ITS, Ms. Lindquist, as measured by student's achievement in algebra. The study is carried out towards the end of software development cycle.

Statement of the Problem

The developer of Ms. Lindquist conducted a simple formative evaluation on the effectiveness of the software by giving 20 students a six-item posttest after 2 hours' use and asserted "Students using Ms. Lindquist did better on the post test". (Heffernan, 2001) This result is not strong enough to show the software's advantage over other computer based or traditional instructional technique. Besides, the sampling and instrument of this research seriously weakened the reliability and validity of the result.

In this proposal, I plan to conduct a more synthesis research on Ms. Lindquist which will provide answer to the question "What's the educational impact of Ms. Lindquist?" The main task of my research will be testing the overall effectiveness of the software to see if it significantly increases the performance in algebra.

Background of the Study

Until recently, little attention has been paid to the evaluation of ITS. Most ITS researchers have concerned themselves with envisioning the potential of ITSs and investigating the implementation issues involved in constructing actual components and systems (Sleeman & Brown, 1982; Wenger, 1987). The literature provides a number of evaluation cases and methods that can be used to evaluate intelligent tutoring systems. However, there are no guidelines available for use of evaluation methods and most of the existing evaluations are empirical endeavors. (Wu, 1996)

Among various types of research, experimental research is most commonly adopted by researchers. There are a variety of experimental research designs such as single group designs, control group designs, and quasi-experimental designs being used by ITS evaluators. (Mark, 1993) But the experimental research enables the researchers to obtain a relationship between a set of treatments and the outcome and as a consequence it is particularly suited to examining the effects of a certain teaching technique. Experimental research has been used in ITS evaluation for VCR (Mark, 1991), Sherlock (Lajoie & Lesgold, 1991), Multimedia ITS (Livergood, 1994), Byzantium ITT (Kinshuk, 2000), and Conceptual Helper (Albacete, et al, 2000).

1. Evaluation of the VCR Tutor

The purpose of this early prototype of ITS was to teach an adult how to program a video cassette recorder (VCR) to automatically record a selected television programs. Actions involved in operating a VCR are procedural, but their execution may depend

upon cognitive knowledge of the device, its behavior, and the relationships between device features and device actions. A cognitive model was embedded in the VCR Tutor which stored all these cognitive knowledge and responded appropriate feedbacks to errors that users made.

In order to evaluate the effectiveness of VCR Tutor, the developer implemented two versions of the software and conducted an experimental research with 20 subjects. The prompting version provided users with several possible procedures (as the next step) to explore at each step, while did not give any feedbacks upon the action performed. The knowledgeable version gave users relative knowledge as feedbacks upon users' errors. The posttest result showed that subjects who used the knowledgeable tutor learned to program a VCR using fewer steps and with fewer errors (also fewer types of errors) than those who used the prompting version.

The researcher concluded that knowledgeable feedback supports learning in the VCR domain more effectively than other types of feedback.

2. Evaluation of the Sherlock Project

The Sherlock Project has developed a computer-based learning-by-doing environment to teach Air Force technicians to efficiently perform the hardest troubleshooting tasks that arise in their jobs, diagnosing faults in a system with thousands of parts. The environment combines a simulation of the system with which they work

every day and a coach that provides advice when they reach impasses while attempting these difficult tasks. Two field tests have shown the system to be highly effective.

Air Force colleagues involved in testing the Sherlock report experimental vs. control effects of two standard deviations, several times the effects usually found with short-term instructional treatments. Further, testing was done with a criterion of real-world performance of the hardest parts of the job and with blind scoring of performances, something quite unusual in educational technology development. Earlier tests of Sherlock established that what is learned is retained (six month retention tests showed losses of no more than 10% of what was learned). The first round of field testing showed that 20-25 hours of Sherlock training produced learning equivalent to about four years of on-the-job experience.

3. A Study of the Effectiveness of Multimedia ITS

In this study of ITS, the author designed his research to isolate the key variables, disallow for the effect of some of those variables, and determine in a precise way what effects are produced by a multimedia delivery system, repetition of instructional and testing material, and an ITS.

In the first phase of this study, comparing test scores of students who were presented material in two delivery systems: 1) hard-copy (printed material), and 2) a computer-based multimedia program, indicated no statistically significant difference in scores. A second part of that phase, comparing test scores and re-test scores of students

studying the computer-based multimedia program, indicated no statistically significant difference in scores.

In the second phase of this study, the multimedia ITS was compared with the two initial delivery systems, which resulted in a statistically significant difference in scores.

4. Evaluation of Byzantium Intelligent Tutoring Tools (ITT)

Although there are some instances of small-scale evaluations that have been completed within a single institution, little work has been reported on large-scale evaluations conducted across several institutions. Kinshuk's ITT evaluation research is the first large-scaled and multi-institutional one. The research evaluated the effectiveness of 3 ITT packages as an alternative to the human-led tutorials.

The research was carried out among students at one university in United Kingdom who studied Capital Investment Appraisal. Students were divided into two parallel groups. One group took traditional teacher's instruction in classroom, while the other group was exposed to ITT tutorials in a computer lab. Group comparison with the help of pre and post tests provided the initial validation of the effectiveness of the ITT, whereas the observations and subjective questionnaire feedback from ITT group validated the interface design adopted in the ITT. Experimental research was adopted since only the overall effectiveness was to be evaluated. The main objective of the research was to determine if the Byzantium project ITTs are an effective alternative to the resource-intensive human-tutor-led tutorials for introductory numeric disciplines.

The statistical analysis showed that the difference between the gains of human tutor based teaching and ITT teaching was not significant while the difference between the gains of various centers is significant.

Another qualitative research, carried out by questionnaire and observations, showed that the overall feelings of the students about the system were quite positive.

The research shows that Byzantium ITT "provides an adequate means of tutoring in the procedural skills that can be employed as an adjunct to the traditional lectures and replace some of the human-led tutorials." The study revealed that the means of gains obtained by the students in the traditional teaching group and ITT teaching group are almost equal. It can be concluded that this ITT is a suitable alternative to human-led tutorials. As there was no significant difference in the performance gain between the students with and without the attributes of previous computer training, confidence in operating computers and enjoyment in using computers, it can be concluded that the Byzantium packages are suitable for all the students learning numeric disciplines.

5. Evaluation of Conceptual Helper ITS

The Conceptual Helper is an ITS designed to coach students through physics homework problem solving of a qualitative nature. Similar to Ms. Lindquist, the design of Conceptual Helper is based on cognitive models. It contains a cognitive model that is capable of correctly solving any problem assigned to the student. Model tracing consists of matching every problem-solving action taken by the student with the steps of the

expert's solution model of the problem being solved. This matching is used as the basis for providing immediate feedback to students as they progress through the problem. The system also has a student model in which each node in the network represents a piece of conceptual knowledge that the student is expected to learn or a misconception that the tutor can help remedy. Each node has a number attached to it that indicates the probability that the student will apply the piece of knowledge when it is applicable. As the student solves a problem, the probabilities are updated according to the actions taken by the student.

Forty-two students taking Introductory Mechanics classes were recruited and randomly divided into a control group and an experimental group. Both groups took a paper-and-pencil pre-test that consisted of 29 qualitative problems. Then both groups solved some problems. The students in the Control Group had their input turned green or red depending on the correctness of the entry. Then, in the case of an incorrect action, the students could ask for help by making a choice from a help menu. If the student asked for more help, they would just be told the correct answer. On the other hand the students in the experimental group received the green/red feedback depending on whether their action was correct but when the input was incorrect the Conceptual Helper intervened as explained above. After the students finished solving the problems with the system they took a post-test which was the same as the pre-test with the exception of a few changes in the cover stories of some problems. Finally the students were asked to complete a questionnaire expressing their evaluation of the system.

The statistical analysis resulted in a significant difference between the experimental group's performance gain and the control group's. The survey research showed that students gave a score of 4 or above to all aspects of the ITS.

Summary of Literature Review

Among above 5 empirical researches, all of them used experimental research with control group as the main method to evaluate the overall effectiveness of an ITS. Research 1, 3, and 5 showed ITS's advantage over menu-driven Computer Based Instruction (CAI). Research 2, 3, 4 showed ITS's advantage over traditional non-computer instruction. Research 4 showed ITS can be an alternative to human tutors.

Research 4 and 5 also conducted a survey research after ITS's intervention, which both got positive feedback from students.

Research 4 tried to find relationship between subjects' pre-treatment attributes and the performance gain, but didn't yield a significant result.

Research Questions

Now, back to my general problem "what's the educational impact of Ms. Lindquist?", based on above empirical methodology and results and the context of my research, my research questions would be:

1. Does Ms. Lindquist significantly increase 7th grade students'performance in writing algebraic expressions from word problems when comparing with traditional CAI? Can it be a competent alternative of human private tutors?

2. How do students like this software?
    - Is the interface easy to use?
    - What difficulty do they have?
    - Are they motivated and interested?
3. Does following student characteristics affect their gains of achievement by using this ITS?
    - Gender
    - Pre-treatment achievement in algebra
    - Pre-treatment computer skills
    - Interest and comfort of using computer

Methods and Procedure

Sampling and Instrumentation:

I would draw a sample of 60 7th grade students from a middle school. The subjects would be randomly and evenly assigned to 3 groups. The control group would be asked to use a menu-driven CAI (only correct/incorrect is responded upon students' answer, and they need to navigate the menu and find relative information); the experimental group 1 would be asked to use Ms. Lindquist (whose feature has been described above); the experimental group 2 would be one-to-one tutored by human tutors. All 3 different types of treatment would be scheduled for 2 hours every day after class, and the whole research would last one week. All 3 groups would be given a pretest and a posttest which contains 20 questions of writing expressions from algebra word questions.

The ITS group would be required to fill out an additional pre-treatment questionnaire which contains following items:

1. gender (male/female)
2. achievement in algebra (percentage of standing in the class)
3. computer skills for each of the following items(5 scales from no experince to proficient)
   a. Operating System (Windows/MacOS)
   b. Office Suite (Word/Excel/Access, etc.)
   c. Programming
   d. Internet surfing and email
   e. Entertainment and Games
4. interest of using computer (5 scales from not interested to very interested)

The ITS group would be again required to fill out a post-treatment questionnaire which contains following items: (each question is to be answered in 5 scales from strongly agree to strongly disagree)

1. The Ms. Lindquist's interface is easy to manipulate
2. I don't have difficulty in using Ms. Lindquist
3. I like Ms. Lindquist
4. I feel achieving more

Data Analysis:

First, since the primary purpose of this research is to evaluate the overall effectiveness of Ms. Lindquist, I would compute and compare the t-value of means of gains of test scores among the 3 groups. A significant difference between control group and ITS group would be expected to show Ms. Lindquist's advantage over CAI. An insignificant difference between ITS group and human tutor group would be expected to show Ms. Lindquist's competency of substituting human tutors.

For the ITS group, correlation coefficients would be computed between each student's score gain and his/her response on each pre-treatment survey questions. This correlation research is to determine if significant correlation exists between student characteristics and their performance gain.

For the ITS group, the information collected from post-treatment survey would be used as a part of formative evaluation for the purpose of improvement of the software.

Limitations

There are several threatens of external validity which weaken the research result's generalizability to the population of all 7^{th} grade students:

1. The cost of research (for computer hardware and software and private tutors) restricts the sample size. It's also hard to draw samples from different institutes.
2. Due to school administration issues, I would probably use convenience sampling, which is not a random one.
3. The high performance may be partially caused by the feeling of novelty. The length of the research is not enough to desensitize this feeling.

The effect of pretest itself would be a threaten of internal validity which may be a factor of score gain.

References

Bloom, BS, (1984). The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring, Educational Researcher 4-16

Duffy, T. M., & Jonassen, D. H. (Eds.). (1992). Constructivism and the technology of instruction: A conversation. Hillsdale NJ: Erlbaum

Elsom-Cook, M. (1987) Intelligent Computer-aided instruction research at the Open University. Technical Report No: 63. Computer-Assisted Learning Research Group, The Open University, Milton Keynes.

Halloun, I.A., & Hestenes, D. (1985). The initial knowledge state of college physics students, American journal of Physics 53 (11) 1043-1055.

Heffernan, N. T (2001) Intelligent Tutoring Systems are Forgotten the Tutor: Adding a Cognitive Model of Human Tutors. Dissertation. Computer Science Department, School of Computer Science, Carnegie Mellon University. Technical Report CMU-CS-01-127

Heller, R. S. (1991). Evaluating software: A review of the options, Computers & Education, 17 (4), 285-291

Lajoie S. P. & Lesgold A. M.(1991): The SHERLOCK experience: An evaluation of a computer-based supported practice environment for electronics trouble-shooting training. Proceedings of the International Conference for Cognitive Science for the Development of Organizations, May 2-4, 1991, Montreal, 56-62.

Mark, M.A (1991) The VCR Tutor: Design and evaluation of an intelligent tutoring system. Master's Thesis, University of Saskatchewan, Saskatoon, Saskatchewan, 1991. Advisor: Prof. J. E. Greer, Department of Computational Science.

Mark M. A. & Greer J. E. (1993): Evaluation methodologies for intelligent tutoring systems. Journal of Artificial Intelligence and Education, 4 (2/3), 129-153.

Michalski R. S. and Chilausky R. L..(1980) Learning by being told and learning from examples: An experimental comparison of the two methodsofknowledge acquisition in the context of developing an expert system for soybean disease diagnosis. International Journal of Policy Analysis and Information Systems, 4(2)

Nwana H.S. (1990). "Intelligent Tutoring Systems: an overview ". Artificial Intelligence Review, 4, p. 251-277.

Norman D. Livergood, Ph.D, (1994). A Study of the Effectiveness of a Multimedia Intelligent Tutoring System , Journal of Educational Technology Systems, Vol. 22(4), 1994, p. 337-344

Major N., Ainsworth S. & Wood D.(1992): REDEEM: Exploiting sumbiosis between psychology and authoring
environments. International Journal of Artificial Intelligence in Education,

Patricia L. Albacete, Kurt A. VanLehn, (2000). Evaluating the Effectiveness of a Cognitive Tutor for Fundamental Physics Concepts, Proceedings of the 22nd Annual Meeting of the Cognitive Science Society

Kinshuk, Ashok Patel, David Russell, (2000). A multi-institutional evaluation of Intelligent Tutoring Tools in Numeric Disciplines, Educational Technology & Society 3(4) 2000

Sleeman, D.H. & Brown, J.S. (Eds.) (1982). Intelligent tutoring systems . New York: Academic Press,

Wu A. K. W.(1996): On the formal evaluation of learning systems. Lecture Notes in Computer Science, 1086, 324-332.