An assessment of the Andes tutor[1]

R. N. Shelby[2], K. G. Schulze[2], D. J. Treacy[2], M. C. Wintersgill[2], K. VanLehn[3] and A. Weinstein[3]
[1]Work supported by ONR, [2] U.S. Naval Academy, [3] LRDC at the University of Pittsburgh

Andes is an intelligent problem-solving tutor for classical physics. It was used in the fall semester of 1999 and 2000 by a group of approximately 150 students. An assessment of its effectiveness was made using: 1) comparative results on a free-response test taken by the Andes group and a control group, 2) portfolios of the students' work and 3) student opinion surveys. The results of the assessment highlight some strengths and weaknesses of Andes.

Introduction

Andes is an intelligent tutoring system in classical physics that is being developed by researchers at the Learning Research and Development Center (LRDC) at the University of Pittsburgh and the United States Naval Academy (USNA). Andes allows students to solve physics problems in an environment that provides visualization, immediate feedback, procedural help, and conceptual help. A description of the Andes system and references can be found at www.press.umich.edu/jep/06-01/schulze.html.

This paper reports an assessment of Andes conducted during the fall semesters of 1999 and 2000 using students in a basic physics course taught at the U. S. Naval Academy. The assessment was done in three parts: 1) free response examination questions, 2) a portfolio of all of the work done by the students on Andes, and 3) a written survey of the participants in the experiment.

Andes was downloaded from a server to the individual computers of the students in the experimental group. Homework assignments in the Andes group were done on the students' individual computers and submitted in class.

In order to understand the results of the trials it is necessary to know some of the features of Andes. A Graphical User Interface (GUI) supports the student in making drawings appropriate to the problem, defining variables to be used, entering relevant equations and obtaining numerical solutions. All of these actions receive immediate feedback; the entry turned green (correct) or red (incorrect). This feature is a particular favorite of the students because it prevents them wasting time by following incorrect paths in their solutions.

There are several other types of help available. If a student is not sure where to start a problem or what to do next in a solution, s/he can ask for a "hint". If a student has taken an incorrect action s/he can ask, "What's wrong?" Both of these requests produce a dialog box with advice. The initial advice is usually fairly broad but relevant to the place in the solution at which the request was made. These dialog boxes contain further options. If a student wishes more specific advice s/he can press one of the hyperlinks in the dialog box "explain further", "how" or "why". There are usually three or four levels of advice below the original with each level becoming more specific. The final level of hints, referred to as a "bottom out" hint, usually tells the student the correct action to take. There are several reasons for including this level of hint but it is certainly open to abuse.

The student is encouraged to make a drawing appropriate to the problem presented (visualization), define the variables to be used in the solution (communication), and enter the appropriate equations in symbolic form. If a complete solution has been accomplished, except for numerical substitution, the student can ask Andes to do the appropriate substitution. This procedure produces solutions that can be evaluated by the instructor. The printed output is organized and contains a drawing, definitions of variables, and symbolic equations. If any of these entries are missing it is easily recognized and can be marked appropriately.

The majority of the problems in Andes during the 1999 and 2000 trials would be classified as exercises. These exercises are used to teach problem solving techniques which are then tested by more difficult problems in Andes. Many of the problems have multiple solution paths, which is one of the most challenging tasks in an artificial intelligence environment.

1. Free Response Examinations

Free response examination questions were used as one of the methods for assessing the effectiveness of Andes as a problem-solving tutor. These questions, two on kinetics and two on Newton's laws, were given to students using Andes and a control group as an hour exam about 6 weeks into the mechanics portion of the first semester of the basic physics course. In addition, two free response questions, one on Newton's laws and one work-energy, were given on the final examination in fall 2000. In each instance the free response questions were graded using a strict rubric that was constructed to enforce effective problem-solving techniques that reflected the goals of the Andes tutor. This rubric stressed defining variables used, using symbolic equations to express concepts, proper expression of vectors, the use of diagrams, and defining coordinate systems where appropriate.

To create a level playing field for the assessment, the grading rubric for the free response questions and the content coverage was publicized to all participants well in advance of the examinations. Also, since previous studies at our institution had shown that a student's grade average (CQPR) and major were the best predictors of performance in the physics course, the Andes and control sections were matched to have similar average CQPR's and major distributions.

There were major differences between the Andes system in 1999 and 2000. The 2000 version had more problems, more extensive coverage of topics and some more difficult problems. In addition, the 2000 version required that all vector variables be defined using a drawing tool with an associated dialog box to specify properties.

The results of the hour exams given in fall 1999 and fall 2000 and the two questions of the fall 2000 final examination are given below.

Table I. Exam#1 - Fall 1999

	Number	Average	S.D.
Andes	173	73.7	13.0
Controls	162	70.4	15.6
Effect size = 0.21 Results of Student's t-Test: t = 2.21 p(null hypothesis) = 0.036			

Table II. Exam#1 - Fall 2000

	Number	Average	S.D.
Andes	140	70.0	13.6
Controls	135	57.1	14.0
Effect size = 0.92 Results of Student's t-Test: t = 7.74 p(null hypothesis) < 0.00001			

Table III. Final Exam - Fall 2000
(2 questions)

	Number	Average	S.D.
Andes	140	75.9	21.3
Controls	135	77.2	21.3
Effect Size = -0.06 Results of Student's t-Test: t = -0.538 p(null hypothesis) = 0.71			

Several observations can be made about the results shown. First, comparing the fall 1999 and fall 2000 Exam #1 results it should be noted that, in an effort to improve discrimination, the 2000 exam was written to be more difficult than the 1999 exam. The expected result of this change was for the average for both groups to decrease and for relative comparisons between Andes and controls to be the only valid comparisons. Using a comparison of the relative results from the two years, about one third of a letter grade higher for the Andes group in 1999 and more than a full letter grade in 2000, it appears that the improvements made in Andes resulted in a more effective tutor. These relative results from both years also indicate that Andes has real promise as an effective tutor. The performance on the two questions on the fall 2000 final exam requires some explanation. First, the problems given were rather easy which resulted in a distribution skewed toward the high end. This ceiling effect made a valid comparison between the two study groups very difficult. Among other possible explanations is the possibility that by the end of the semester the control students had finally become convinced that they would have to use accepted techniques for presenting physics problem solutions.

2. Log Portfolios

A second form of assessment which can be applied to Andes is a portfolio method using the students' log files. Every keystroke a student makes, whether it is correct or incorrect, is recorded in a log file. At the end of a session, when the student exits from Andes, this file is uploaded to a local server and saved. The instructors can access these log files.

These files can also be used to help a student who is having difficulty. The student and the instructor can sit down together and replay the log file. At a place where the student was having difficulty the replay can be stopped and a dialog established to help the student understand a concept or procedure.

The log files represent a portfolio of all the work each student did during the portion of the semester covered by Andes. This produced 136 individual student portfolios during the fall of 2000. A Perl program was used to read the log files and collate the data. The final form of the data was an Excel workbook listing the significant parameters. Requests for various forms of help available from Andes are documented with unique codes. We examined the types of help requested and their frequency.

The overall picture, which came from this assessment, is that all forms of help were used by a complete cross section of students. On the average a student with a high CQPR was just as likely to ask for a hint as a student with a low CQPR. This data was examined by quintile for the Andes experimental group. It was sorted both by CQPR to determine the average number of hints per student and per problem as well as being sorted by the total number of requests for help and then determining an average CQPR.

Some interesting results are: the average student worked 49.5 out of 60 problems assigned; the average time per problem was 22.8 minutes; the average number of hints and equation corrections requested per student was 270, which translates to 5.4 requests per problem; and the average number of explanations requested was 430, which is about 8.6 per problem.

The average time per problem is an accurate measure of the time that the student spent actively working on the problem. If a student takes more than 10 minutes between two actions, this time is considered as "time wasted" and is not included in the working time.

One of the significant features of this analysis is that it allows us to determine whether problems were worked independently or if they were copied electronically. For problems, which were copied, the logs show that it takes approximately 7 seconds to read a solution done by another student. All solutions of this nature are excluded from the data.

Another feature that can be addressed with this data is the question of abuse of the help system. In model tracing tutors it is possible to ask for a hint and keep asking for further explanations down to the lowest level possible. At this lowest level the hint is usually more explicit than an instructor would like. We have looked at the average ratio of explanations to initial requests and found it to be about 1.6. Since there are three levels of explanations in every hint this says that on the average the students are not abusing the help system. The logs of the students who are at least two standard deviations above this ratio were examined. Only 4 out of 136 log files showed an indication of help abuse.

The log file portfolio told us that the average student was reasonably serious about doing the assigned homework, s/he used the help system regularly, and the requests for help were fairly uniform across the entire spectrum of CQPRs.

3. Student Surveys

Student reaction to the use of Andes in fall 2000 was recorded by use of a questionnaire and from comments made on end-of-course evaluation forms. The questionnaire data were examined as one set for all students in the Andes group and in five groups (1 through 5) determined by the student's indication of what fraction (< 0.25 to >0.8) of the Andes problems s/he had completed independently. On questions like "How much did you enjoy using Andes vs. paper and pencil", "Did you learn more using Andes?" "Did you do more exercises using Andes than you would have done using paper and pencil?" the average response was neutral. However, the grouped data indicated that the more independent work a student did using Andes the more favorably they viewed the program. There was general agreement that doing a problem using Andes did take somewhat more time than using a standard method. There was also agreement that immediate feedback was useful. Using tools and dialog boxes to draw diagrams and define vectors, and having Andes solve algebraic equations and substitute values into equations were viewed as being positive aspects of the tutor.

Finally, the students were asked if they would have used Andes after zero, two, six and ten weeks, if its use had not been required. For the complete Andes group the indication of average voluntary usage dropped from 48% at the beginning of the semester to 28% at the end of the semester. For students who did more than 80% of the Andes problems independently the data show that the voluntary usage would have remained relatively constant at about 45%.

There are several indicators in the data that tell us that help must be improved and that we must have more effective methods of giving instructions for use of the system.

Course evaluation forms completed by students at the end of the semester were scanned to locate any unsolicited comments about Andes. This search identified 113 forms with comments, which were classified on a qualitative scale. The comments covered a wide range of possibilities from "Andes was a waste of time." and "Homework is much easier using pencil and paper." to "Andes was a pain to get done, but it definitely helped to understand the material by forcing certain method use and thought processes." A general theme in the comments was that the students resented the precision of the notation required in the solutions, but many slowly came to appreciate its importance. Table IV summarizes these comments.

Table IV. Student comments

Positive or very positive	50%
Neutral	12%
Negative or very negative	38%

An overview of this assessment shows that: 1) that Andes shows promise as a coached problem-solving environment for basic physics; and 2) improvement is necessary in the quality of the help and the instruction given in the use of the system. These improvements are being incorporated into the trial for the fall of 2001.