The Andes Physics Tutoring System: Five Years of Evaluations

Kurt VANLEHN[1], Collin Lynch[1], Kay Schulze[2], Joel A. Shapiro[3], Robert Shelby[4], Linwood Taylor[1], Don Treacy[4], Anders Weinstein[1], and Mary Wintersgill[4]
[1] *LRDC, University of Pittsburgh, Pittsburgh, PA, USA*
[2] *Computer Science Dept., US Naval Academy, Annapolis, MD, USA*
[3] *Dept. of Physics and Astronomy, Rutgers University, Piscataway, NJ, USA*
[4] *Physics Department, US Naval Academy, Annapolis, MD, USA*

Abstract. Andes is a mature intelligent tutoring system that has helped hundreds of students improve their learning of university physics. It replaces pencil and paper problem solving homework. Students continue to attend the same lectures, labs and recitations. Five years of experimentation at the United States Naval Academy indicates that it significantly improves student learning. This report describes the evaluations and what was learned from them.

1 Introduction

Although many students have personal computers now and many effective tutoring systems have been developed, few academic courses include tutoring systems. A major point of resistance seems to be that instructors care deeply about the content of their courses, even down to the finest details. Most instructors are not completely happy with their textbooks; adopting a tutoring system means accommodating even more details that they cannot change.

Three solutions to this problem have been pursued. One is to include instructors in the development process. This lets them get the details exactly how they want them, but this solution does not scale well. A second solution is to include the tutoring system as part of a broader reform with significant appeal to instructors. For instance, the well-know Cognitive Tutors (www.carnegielearning.com) are packaged with an empirically grounded, NCTM-compliant mathematics curriculum, textbook and professional development program. A third solution is to replace grading, a task that many instructors would rather delegate anyway. This is the solution discussed here.

The rapid growth in web-based homework (WBH) grading services, especially for college courses, indicates that instructors are quite willing to delegate grading to technology. In physics, the task domain discussed here, popular WBH services include WebAssign (www.webassign.com), CAPA (www.lon-capa.org/index.html) and Mastering Physics (www.masteringphysics.com). Ideally, instructors still chose their favorite problems from their favorite textbooks, and they may still use innovative interactive instruction during classes and labs. [1] All that changes is that students enter their homework answers on-line, and the system provides immediate feedback on the answer. If the answer is incorrect, the student may receive a hint and may get another chance to derive the answer. Student homework scores are reported electronically to the instructor.

Although WBH saves instructors time, the impact on student learning is unclear. WBH's immediate feedback might increases learning relative to paper-and-pencil homework, or it might increase guessing and thus hurt learning. Although most studies merely report correlations between WBH usage and learning gains, 3 studies of physics instruction have compared learning gains of WBH to those of paper-and-pencil homework (PPH). In the first study, [2] one of 3 classes showed more learning with WBH than PPH. Unfortunately, PPH homework was not collected and graded, but WBH was. It could be that the WBH students did more homework, which in turn caused more learning. In the other studies, [3, 4] PPH problem solutions were submitted and graded, so students in the two conditions solved the roughly the same problems for the same stakes. Despite a large number of students and an impressive battery of assessments, none of the measures showed a difference between PPH students and WBH students. In short, WBH appears to neither benefit nor harm students' learning compared to PPH.

The main goal of the Andes project is to develop a system that is like WBH in that it replaces only the PPH of a course, and yet it increases student learning. Given the null results of the WBH studies, this appears to be a tall challenge. This paper discusses Andes only briefly—see [5] for details. It focuses on the evaluations that test whether Andes increases student learning compared to PPH.

2 The function and behavior of Andes

In order to make Andes' user interface easy to learn, it is as much like pencil and paper as possible. A typical physics problem and its solution on the Andes screen are shown in Figure 1. Students read the problem (top of the upper left window), draw vectors and coordinate axes (bottom of the upper left window), define variables (upper right window) and enter equations (lower right window). These are actions that they do when solving physics problems with pencil and paper.

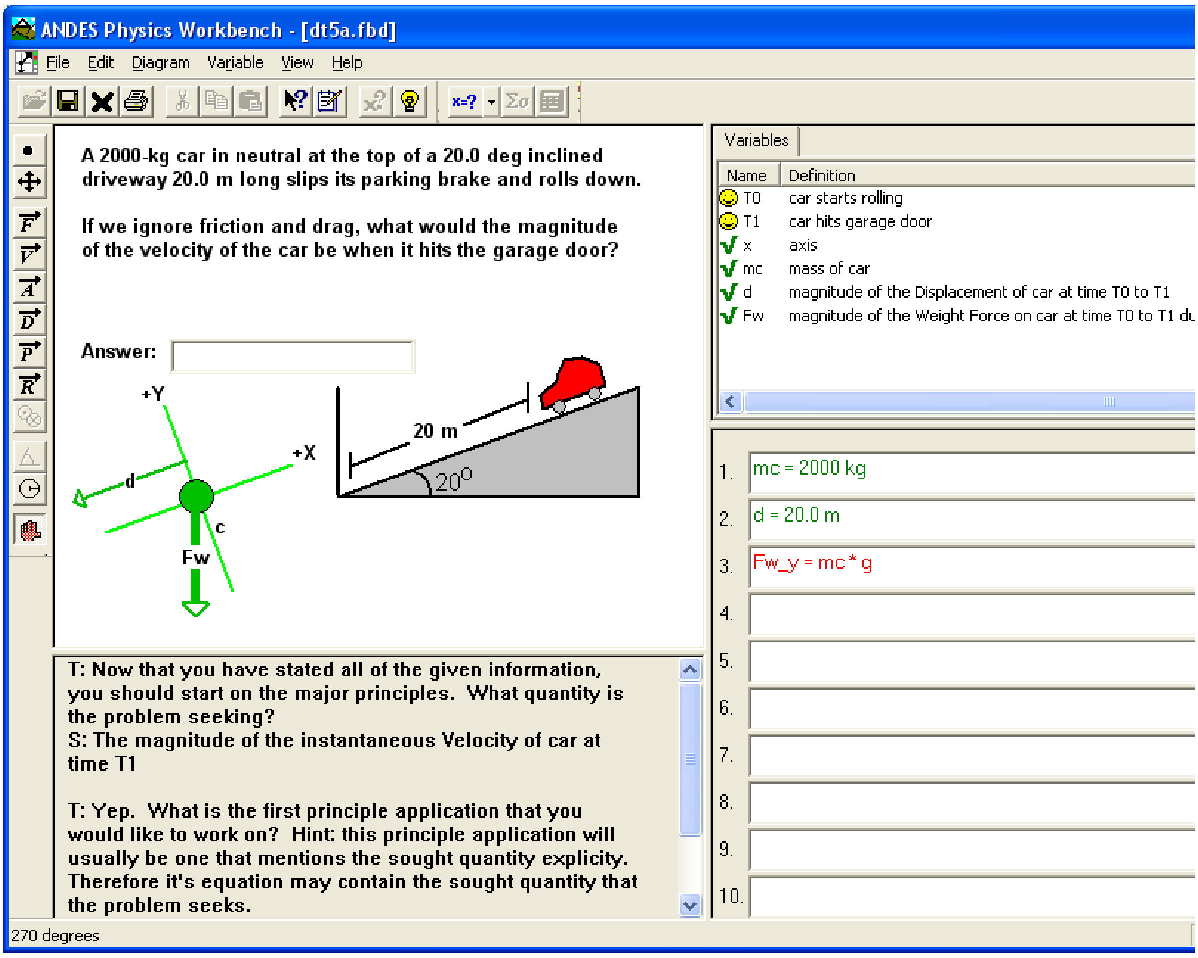
Unlike PPH, as soon as an action is done, Andes gives immediate feedback. Entries are colored green if they are correct and red if they are incorrect. In Figure 1, all the entries are green except for equation 3, which is red.

Also unlike PPH, variables and vectors must be defined before being used. Vectors and other graphical objects are first drawn by clicking on the tool bar on the left edge of Figure 1, then drawing the object using the mouse, then filling out a dialogue box. Filling out these dialogue boxes forces students to precisely define the semantics of variables and vectors. For instance, when defining a force, the student uses menus to select two objects: the object that the force acts on and the object the force is due to.

Andes includes a mathematics package. When students click on the button labeled "x=?" Andes asks them what variable they want to solve for, then it tries to solve the system of equations that the student has entered. If it succeeds, it enters an equation of the form <variable> = <value>. Although physics students routinely use powerful hand calculators, Andes' built-in solver is more convenient and avoids calculator typos.

Andes provides three kinds of help:

- Andes pops up an error messages whenever the error is probably due to lack of attention rather than lack of knowledge. Typical slips are leaving a blank entry in a dialogue box, using an undefined variable in an equation (which is usually caused by a typo), or leaving off the units of a dimensional number. When an error is not recognized as a slip, Andes merely colors the entry red.
- Students can request help on a red entry by selecting it and clicking on a help button. Since the student is essentially asking, "what's wrong with that?" we call this *What's Wrong Help*.


ANDES Physics Workbench - [dt5a.fbd]
File Edit Diagram Variable View Help
x=?
Variables
+Y
+X
d
c
Fw
20 m
20°

A 2000-kg car in neutral at the top of a 20.0 deg inclined driveway 20.0 m long slips its parking brake and rolls down.

If we ignore friction and drag, what would the magnitude of the velocity of the car be when it hits the garage door?

Answer:

Name	Definition
☺ T0	car starts rolling
☺ T1	car hits garage door
✓ x	axis
✓ mc	mass of car
✓ d	magnitude of the Displacement of car at time T0 to T1
✓ Fw	magnitude of the Weight Force on car at time T0 to T1 du

1. mc = 2000 kg
2. d = 20.0 m
3. Fw_y = mc * g
4.
5.
6.
7.
8.
9.
10.

T: Now that you have stated all of the given information, you should start on the major principles. What quantity is the problem seeking?
S: The magnitude of the instantaneous Velocity of car at time T1

T: Yep. What is the first principle application that you would like to work on? Hint: this principle application will usually be one that mentions the sought quantity explicity. Therefore it's equation may contain the sought quantity that the problem seeks.

270 degrees

Figure 1: The Andes screen (truncated on the right)

- If students are not sure what to do next, they can click on a button that will give them a hint. This is called *Next Step Help*.

What's Wrong Help and Next Step Help generate a hint sequence that usually has three hints: a pointing hint, a teaching hint and a bottom-out hint. As an illustration, suppose a student who is solving Figure 1 has asked for What's Wrong Help on the incorrect equation *Fw_x = -Fw*cos(20 deg)*. The first hint, which is a pointing hint, is "Check your trigonometry." It directs the students' attention to the location of the error, facilitating self-repair and learning. [6, 7] If the student clicks on "Explain more", Andes gives a teaching hint, namely:

> If you are trying to calculate the component of a vector along an axis, here is a general formula that will always work: Let θV be the angle as you move counterclockwise from the horizontal to the vector. Let θx be the rotation of the x-axis from the horizontal. (θV and θx appear in the Variables window.) Then: V_x = V*cos(θV-θx) and V_y = V*sin(θV-θx).

We try to keep teaching hints as short as possible, because students tend not to read long hints. [8, 9] In other work, we have tried replacing the teaching hints with either multimedia [10, 11]or natural language dialogues. [12] These more elaborate teaching hints significantly increased learning, albeit in laboratory settings.

If the student again clicks on "Explain more," Andes gives the bottom-out hint, "Replace cos(20 deg) with sin(20 deg)." This tells the student exactly what to do.

Andes sometimes cannot infer what the student is trying to do, so it must ask before it can give help. An example is shown in Figure 1. The student has just asked for Next Step Help and Andes has asked, "What quantity is the problem seeking?" Andes pops up a menu or a dialogue box for students to supply answers to such questions. The students' answer is echoed in the lower left window.

In most other respects, Andes is like WBH. Instructors assign problems via email. Students submit their solutions via the web. Instructors access student solutions via a spreadsheet-like gradebook. They can accept Andes' scores for the problems or do their own scoring, and so on.

3 Evaluations

Andes was evaluated in the U.S. Naval Academy's introductory physics class every fall semester from 1999 to 2003. This section describes the 5 evaluations and their results.

Andes was used as part of the normal Academy physics course. The course has multiple sections, each taught by a different instructor. Students in all sections take the same final exam and use the same textbook but different instructors assign different homework problems and give different hour exams, where hour exams are in-class exams given approximately monthly. In sections taught by the authors (Shelby, Treacy and Wintersgill), students were encouraged to do their homework on Andes. Each year, the Andes instructors recruited some of their colleagues' sections as Controls. Students in the Control sections did the same hour exams as students in the Andes section.

Control sections did homework problems that were similar but not identical to the ones solved by Andes students. The Control instructors reported that they required students to hand in their homework, and credit was given based on effort displayed. Early in the semester, instructors marked the homework carefully in order to stress that the students should write proper derivations, including drawing coordinate systems, vectors, etc. Later in the semester, homework was graded lightly, but instructors' marks continued the emphasis on proper derivations. In some classes, instructors gave a weekly quiz consisting of one of the problems from the preceding homework assignment. All these practices encouraged Control students to both do the assignments carefully and to study the solutions that the instructor handed out.

The same final exams were given to all students in all sections. The final exams comprised approximately 50 multiple choice problems to be solved in 3 hours. The hour exams had approximately 4 problems to be solved in 1 hour. Thus, the final exam questions tended to be less complex (3 or 4 minutes each) than the hour exam questions (15 minutes each). On the final exam, students just entered the answer, while on the hour exams, students showed all their work to derive an answer. The hour exam results will be reported first.

3.1 Hour exam results

Table 1 shows the hour exam results for all 5 years. It presents the mean score (out of 100) over all problems on one or more exams per year. In all years, the Andes students scored reliably higher than the Control students with moderately high effect sizes, where effect size defined as (Andes_mean – Control_mean)/Control_standard_deviation. The

Table 1: Hour exam results

Year	1999	2000	2001	2002	2003	Overall
Andes students	173	140	129	93	93	455
Control students	162	135	44	53	44	276
Andes mean (SD)	73.7 (13.0)	70.0 (13.6)	71.8 (14.3)	68.2 (13.4)	71.5 (14.2)	0.22 (0.95)
Control mean (SD)	70.4 (15.6)	57.1 (19.0)	64.4 (13.1)	62.1 (13.7)	61.7 (16.3)	-0.37 (0.96)
P(Andes= Control)	0.036	< .0001	.003	0.005	0.0005	<.0001
Effect size	0.21	0.92	0.52	0.44	0.60	0.61

1999 evaluation had a lower effect size, probably because Andes had few physics problems and some bugs, thus discouraging students from using it. It should probably not be considered representative of Andes' effects, and will be excluded from other analyses in this section.

In order to calculate overall results (rightmost column of Table 1), it was necessary to normalize the exam scores because the exams had different grand means in different years (the grand mean includes all students who took the exam). Each student's exam score was converted to a z-score, where z_score = (score – grand_mean) ÷ grand_standard_deviation. The z-scores from years 2000 through 2003 were aggregated. The overall effect size was 0.61.

The physics instructors recognize that the point of solving physics problems is not to get the right answers but to understand the reasoning involved, so they used a grading rubric for the hour exams that scored the students' work in addition to their answers. In particular, 4 subscores were defined (weights in the total score are shown in parentheses):

- *Drawings:* Did the student draw the appropriate vectors, axes and bodies? (30%)
- *Variable definitions:* Did the student use standard variable names or provide definitions for non-standard names? (20%)
- *Equations:* Did the student display major principle applications by writing their equations without algebraic substitutions and otherwise using symbolic equations correctly? (40%)
- *Answers:* Did the student calculate the correct number with proper units? (10%)

Andes was designed to increase student conceptual understanding, so we would expect it to have more impact on the more conceptual subscores, namely the first 3. Table 2 shows the effect sizes, with p-values from two-tailed t-tests shown in parentheses. Results are not available for 2001. Two hour exams are available for 2002, so their results are shown separately.

There is a clear pattern: The skills that Andes addressed most directly were the ones on which the Andes students scored higher than the Control students. For two subscores, Drawing and Variable definitions, the Andes students scored significantly higher then the Control students in every year. These are the problem solving practices that Andes requires students to follow.

The third subscore, Equations, can also be considered a measure of conceptual understanding. However, prior to 2003, Andes was incapable of discriminating between good and poor usage of equations, so it is not surprising that the Andes and Control students tied on the Equations subscore in years 2000 and 2002. In 2003, Andes gave students warnings and points off on their problem scores if their first use of a major principle was combined algebraically with other equations. Although Andes could have required students to obey this problem solving practice, it only suggested it. This may explain why the Andes students still did no better than the Control students on the Equations subscore in 2003.

The Answers subscore was the same for both groups of students for all years even though the Andes students produced better drawings and variable definitions on those tests. This suggests that the probability of getting a correct answer depends strongly on other skills, such as algebraic manipulation, that are not measured by the more conceptual subscores and not emphasized by Andes. The tied Answer subscores suggest that the

Table 2: Hour exam effect sizes broken down by subscore

Year	**2000**	**2002a**	**2002b**	**2003**	**Average**
Drawings	1.82 (<.001)	0.49 (.003)	0.83 (<.001)	1.72 (<.001)	1.21
Variable definitions	0.88 (<.001)	0.42 (.009)	0.36 (.026)	1.11 (<.001)	0.69
Equations	0.20 (.136)	0.12 (.475)	0.30 (.073)	-0.17 (.350)	0.11
Answers	-0.10 (.461)	-0.09 (.585)	0.06 (.727)	-0.20 (.154)	-0.08

Andes students' use of the equation solving tool did not seem to hurt their algebraic manipulation on the hour exams.

3.2 Final Exam scores

A final exam covers the whole course, but Andes does not. However, its coverage has steadily increased over the years. In 2003, Andes covered 70% of the homework problems in the course. This section reports an analysis of the 2003 final exam data.

In this physics course, engineering and science majors tend to score higher on the final exam than other majors. Unfortunately, there were reliably more engineers among the Andes students than the non-Andes students ($p < .0001$, 3x2 Chi-squared test). Thus, for each group of majors, we regressed the final exam scores against the students' GPAs. (Of the 931 students, we discarded scores from 19 students with unclassifiable majors or extremely low scores). This yielded three statistically reliable linear models, one for each type of major. For each student, we subtracted the exam score predicted by the linear model from the student's actual score. This residual score represents how much better or worse this student scored compared to the score predicted solely on the basis of their GPA and their major. That is, the residual score factors out the students' general competence. The logic is the same as that used with an ANCOVA, with GPA and major serving as covariates instead of pre-test scores. (This kind of statistical compensation was unnecessary in our analysis of the hour exams, because the distributions of majors and student GPAs did not differ across conditions in any year.)

Using these residual scores, we evaluated Andes' impact on students in each of the 3 groups of majors. As Table 3 indicates, the residual scores of the engineering and science majors were not statistically different with Andes than with paper homework. However, the other majors did learn more with Andes than with paper homework ($p=0.013$; effect size = 0.52). Over all students, the mean residual scores for Andes students was higher than for non-Andes students ($p=0.028$; effect size = 0.25).

As though we were gratified to see that Andes students learned more than non-Andes students, we were not surprised that that Andes had little effect on the learning of the engineering and science majors, for two reasons. (1) In many studies, instructional manipulations tend to affect only the less competent students' learning, because highly competent students can usually learn equally well from the experimental and the control instruction [13]. (2) The engineering majors were concurrently taking a course on Statics, which has very similar content to the physics courses. This dilutes the effect of Andes, since it affected only their physics homework and not their Statics homework.

3.3 Comparing Andes to the "benchmark" system

Next we compare our results to results from one of the few large-scaled, controlled field studies of intelligent tutoring systems in the open literature, namely, the evaluation of a combination of an intelligent tutoring system (PAT) and a novel curriculum (PUMP), which is now distributed by Carnegie Learning as the Algebra I Cognitive Tutor. The evaluation was conducted by Koedinger et al. [13]. It is arguably the benchmark against

Table 3: Residual scores on the 2003 final exam

	Engineers	Scientists	Others	All
Andes students	55	9	25	89
Non-Andes students	278	142	403	823
Andes students mean (SD)	0.74 (5.51)	1.03 (3.12)	2.91 (6.41)	1.38 (5.65)
Non-Andes students mean (SD)	0.00 (5.39)	0.00 (5.79)	0.00 (5.64)	0.00 (5.58)
p(Andes=non-Andes)	0.357	0.621	0.013	0.028
Effect size	0.223	0.177	0.52	0.25

which all other tutoring systems should be compared.

Koedinger et al. used both experimenter-defined and standardized tests. Using the experimenter-designed tests, they found effect sizes of 1.2 and 0.7. In our evaluation, the closest matching measures are the Diagram and Variables components of the hour exams, which tap the conceptual skills most directly taught by Andes. Surprisingly, these assessments had exactly the same effect sizes as the Koedinger et al. tests: Diagrams: effect size 1.21; Variables: effect size 0.69.

Koedinger et al. found smaller effect sizes, 0.3, when using multiple-choice standardized tests. The standardized tests most closely match our multiple-choice final exam, where Andes students scored marginally higher than non-Andes students with an effect size of 0.25.

Thus, the Andes evaluations and the Koedinger et al. evaluations have remarkably similar tests and effect sizes. They both have impressive 1.2 and 0.7 effect sizes for conceptual, experimenter-designed tests, and lower effect sizes on standardized, answer-only tests.

The Andes evaluations differed from the Koedinger et al. evaluation in a crucial way. The Andes evaluations manipulated only the way that students did their homework—on Andes vs. on paper. The evaluation of the Pittsburgh Algebra Tutor (PAT) was also an evaluation of the Pittsburgh Urban Mathematics Project curriculum (PUMP), which focused on analysis of real world situations and the use of computational tools such as spreadsheets and graphers. It is not clear how much gain was due to the tutoring system and how much was due to the new curriculum. In our evaluation, the curriculum was not reformed. The gains in our evaluation are a better measure of the power of intelligent tutoring systems *per se*. This is good news for the whole field of intelligent tutoring systems.

4 Conclusions and future work

It appears that we have succeeded in finding a way to use intelligent tutoring systems to help students learn while replacing only their paper-and-pencil homework. Moreover, Andes is probably more effective than existing WBH services, such as WebAssign, CAPA and Mastering Physics. The existing evaluations, which were reviewed in the introduction, suggest that WBH is no more effective than paper-and-pencil homework (PPH), whereas Andes is significantly more effective than PPH. The effect sizes for the open response and multiple choice exams are 0.61 and 0.25, respectively. To be certain that Andes is more effective than WBH, however, one should compare it directly to one of these systems.

We have also shown that Andes' benefits are similar in size to those of the "benchmark" intelligent tutoring system developed by Anderson, Corbett and Koedinger and now distributed by Carnegie Learning. However, Andes' benefits were achieved without attempting to reform the content of the course.

For the immediate future, we have three goals. The first is to help people all over the world use Andes as the U.S. Naval Academy has done, as a homework helper for their courses. Please see www.andes.pitt.edu if you are interested, and please view the training video before trying to use the system.

The second goal is to develop a self-paced, open physics course based on Andes based on mastery learning. We are currently looking for instructors who are interested in developing such a self-paced physics course with us. Please write us if you are interested.

Lastly, the Pittsburgh Science of Learning Center (www.learnlab.org) uses Andes in its physics LearnLab course. A LearnLab course is a regular course that has been heavily instrumented so that investigators can test hypotheses with the same rigor as they would obtain in the laboratory, but with the added ecological validity of a field setting.

5 Acknowledgements

This research was supported by the Cognitive Sciences Program of the Office of Naval Research under grants N00019-03-1-0017 and ONR N00014-96-1-0260, and by NSF under grant SBE-0354420. We gratefully acknowledge the Andes Alumni: Drs. Patricia Albacete, Cristina Conati, Abigail Gertner, Zhendong Niu, Charles Murray, Stephanie Siler, and Ms. Ellen Dugan

6 References

[1] R. R. Hake, "Interactive-engagement vs. traditional methods: A six-thousand student survey of mechanics test data for introductory physics students," *American Journal of Physics*, vol. 66, pp. 64-74, 1998.

[2] R. J. Dufresne, J. P. Mestre, D. M. Hart, and K. A. Rath, "The effect of web-based homework on test performance in large enrollment introductory physics courses," *Journal of Computers in Mathematics and Science Teaching*, vol. 21, pp. 229-251, 2002.

[3] S. W. Bonham, D. L. Deardorff, and R. J. Beichner, "Comparison of student performance using web and paper-based homework in college-level physics," *Journal of Research in Science Teaching*, vol. 40, pp. 1050-1071, 2003.

[4] A. M. Pascarella, "CAPA (Computer-Assisted Personalized Assignments) in a Large University Setting," in *Education*. Boulder, CO: University of Colorado, 2002.

[5] K. Vanlehn, C. Lynch, K. Schultz, J. A. Shapiro, R. H. Shelby, L. Taylor, D. J. Treacy, A. Weinstein, and M. C. Wintersgill, "The Andes physics tutoring system: Lessons learned," *International Journal of Artificial Intelligence and Education*, in press.

[6] G. Hume, J. Michael, A. Rovick, and M. Evens, "Hinting as a tactic in one-on-one tutoring," *Journal of the Learning Sciences*, vol. 5, pp. 23-49, 1996.

[7] D. C. Merrill, B. J. Reiser, M. Ranney, and J. G. Trafton, "Effective tutoring techniques: A comparison of human tutors and intelligent tutoring systems," *The Journal of the Learning Sciences*, vol. 2, pp. 277-306, 1992.

[8] J. R. Anderson, A. T. Corbett, K. R. Koedinger, and R. Pelletier, "Cognitive Tutors: Lessons Learned," *The Journal of the Learning Sciences*, vol. 4, pp. 167-207, 1995.

[9] J.-F. Nicaud, D. Bouhineau, C. Varlet, and A. Nguyen-Xuan, "Towards a product for teaching formal algebra," in *Artificial Intelligence in Education*, S. P. Lajoie and M. Vivet, Eds. Amsterdam: IOS Press, 1999, pp. 207-214.

[10] P. L. Albacete and K. VanLehn, "Evaluation the effectiveness of a cognitive tutor for fundamental physics concepts," in *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society*, L. R. Gleitman and A. K. Joshi, Eds. Mahwah, NJ: Erlbaum, 2000, pp. 25-30.

[11] P. L. Albacete and K. VanLehn, "The Conceptual Helper: An intelligent tutoring system for teaching fundamental physics concepts," in *Intelligent Tutoring Systems: 5th International Conference, ITS 2000*, G. Gauthier, C. Frasson, and K. VanLehn, Eds. Berlin: Springer, 2000, pp. 564-573.

[12] C. P. Rose, A. Roque, D. Bhembe, and K. VanLehn, "A hybrid language understanding approach for robust selection of tutoring goals," in *Intelligent Tutoring Systems, 2002: 6th International Conference*, S. A. Cerri, G. Gouarderes, and F. Paraguacu, Eds. Berlin: Springer, 2002, pp. 552-561.

[13] L. J. Cronback and R. E. Snow, *Aptitudes and instructional methods: A handbook for research on interactions.* New York: Irvington, 1977.