Andes3: An intelligent homework tutor for a variety of high-school physics courses

Andes2 is an existing intelligent tutoring system for physics. It was developed for the United States Naval Academy (USNA). Extensive evaluations indicated that it was successful in this particularly homogeneous context. In this proposal, we will first identify 4 major problems of extending Andes2 to multiple high-school instructional contexts. Then we will describe our proposed solutions. To implement these solutions, we will develop a new version of the system, called Andes3**,** over several cycles of implementation and formative evaluation during a 3 year period. See Table 1. A summative evaluation of Andes3 will be conducted in a Goal 3 extension of this project.

We believe that the 4 problems faced by Andes2 are common to other educational technologies that are developed initially in a homogenous setting, and that our proposed solutions to these problems may work for some of the other technologies as well. Thus, this project contributes both to physics education research and to educational technology research.

1. Significance

A report from the National Science Board (2003) recommends that in order to ensure the country's capacity in science and engineering in an increasingly competitive and changing global labor market, "The Federal Government and its agencies must step forward to ensure the adequacy of the US science and engineering workforce. All stakeholders must mobilize and initiate efforts that increase the number of US citizens pursuing science and engineering studies and careers." Introductory physics is on the critical path to this important goal, as it is a requirement for virtually all advanced training in science and engineering. However, despite the fact that, fully one-third of recent high-school graduates have taken physics (Hehn & Neuschatz, 2006), "The net result is that almost 30 percent of high school graduates enter college unprepared for first-year coursework or arrive at the workplace without the mathematical, scientific, and technical skills that employers require" (National Science Board, 2007). Clearly, high school physics courses need to be improved.

Fortunately, cognitive science has often used physics as the task domain in its studies of fundamental issues in learning and cognition. For instance, introductory, college-level physics was the task domain for seminal work in:

- expert-novice differences (Chi, Feltovich, & Glaser, 1981; Larkin, McDermott, D. P. Simon, & H. A. Simon, 1980; Priest & Lindsay, 1992),
- self-explanation of examples (Chi, Bassok, Lewis, P. Reimann, & Glaser, 1989; Chi & Kurt VanLehn, 1991),
- analogical problem solving (Larkin, Reif, Carbonell, & Gugliotta, 1988; K. VanLehn & R. M Jones, 1993; Bassok & Holyoak, 1995),
- human tutoring (Chi, S. Siler, & Jeong, 2004; K. VanLehn, S. Siler, Murray, Yamauchi, & Baggett, 2003; Chi, Roy, & Hausmann, 2008; VanLehn et al., 2005), and
- collaborative problem solving (Chi et al., 2008; Robert G. M. Hausmann, B. van de Sande, C. van de Sande, & VanLehn, 2008; Hausmann & Chi, 2002).

On the basis of this work, several cognitive models of physics learning have been developed (Elio & Scharf, 1990; Reimann, Wichmann, & Schult, 1993; VanLehn, Jones, & Chi, 1992) leading to a consensus view of how physics expertise is acquired (VanLehn & van de Sande). The Andes project was built on this solid foundation of physics cognitive science (VanLehn et al., 2005).

Andes is neither a curriculum nor a replacement for teachers. It is more like an electronic workbook. That is, instructors periodically assign a selection of problems from the 500+ problems available in Andes to be done as homework. Students solve the problems with immediate feedback and help on request from Andes. Andes grades the solutions and submits them electronically to the instructor.

Evaluations showed that Andes was highly effective in the USNA context, and it remains in use there. However, as we have tried to increase the set of schools that use Andes2, we have become aware of the heterogeneity of other contexts. For instance, all sophomores in the USNA take the same physics course, whereas, at the high school level, there are 4 commonly offered courses: conceptual physics, regular physics, honors physics, and AP physics (Hehn & Neuschatz, 2006), as well as "physics first" courses for younger students (Ewald, J. B. Hickman, P. Hickman, & Myers, 2005). Andes2 has now been used successfully for over a year in three of these courses (regular, honors, and AP) by two instructors at Watchung Hills Regional High School in New Jersey. Five instructors are planning to use Andes2 in these classes during the 2008-2009 school year. The Watchung Hills experience has helped us understand Andes2's limitations and how it must be extended.

Our understanding of the homogeneity issue was broadened as we attended national and regional conferences for physics teachers, where we gave talks, presented posters and had many informal demo/discussion sessions with individual high school and college physics instructors. We have cataloged the reasons that instructors who liked the Andes concept were not able to use it for their classes.

Based on the relatively successful Watchung Hills experience as well as discussions with non-adopting instructors, we believe that four major types of extensions are needed in order to meet the needs of our target physics classes (Regular, Honors, & AP):

1. *Content*. Although Andes2 has over 500 physics problems covering almost all topics of introductory physics, instructors often want to assign problems that Andes does not yet have. Moreover, they often want types of tasks that Andes2 does not support at all, such as free-hand drawing of time-motion graphs. This is particularly true of the Regular Physics classes, which often use graphical and text-based approaches instead rather than algebraic methods.
2. *Scaffolding*. Andes3 needs to provide richer scaffolding in order to make its activities accessible and effective for all students. The USNA is a highly selective institution with highly motivated students, and physics is taught in the sophomore year. High school students are younger, often less motivated and often less mathematically prepared.
3. *Customization*. Many high school instructors are active researchers in that they try a variety of methods for improving their students' learning. Andes3 needs to let such instructors control and customize its scaffolding, hinting behaviors, and content.

4. *Usability.* High school students and instructors do their work on a large variety of computers and networks. Andes3 needs to run on all of them. Andes2 has a complex user interface that takes at least a half-hour of interface-specific training to learn. Andes3 needs a user interface that students can learn without any specific pre-training at all.
5. *Support.* In order to maintain a high level of user support while substantially increasing the number of users, we plan to develop communities of users and developers.

Thus, the significance of the proposed research is two-fold. As our first contribution, we will extend an existing technology, Andes2, that is known to be effective in a single, relatively homogenous setting (the USNA), to a wide variety of high school settings, including multiple course types (regular, honors, & AP), multiple schools, multiple instructors, multiple instructional objectives, and many different kinds of students. By bringing the benefits of this proven technology to as many students as possible, the nation moves one step closer to training a scientifically literate workforce.

Our second contribution will be documenting in the educational technology literature our progress on extending Andes to more heterogeneous settings. Many promising technologies are developed in a homogenous setting and fail to go beyond them. We hope to show others how such extensions can be accomplished. Thus, other critical topics besides physics may benefit from the proposed research.

In the remainder of this section, we amplify the themes introduced above. In the first subsection below ("Current state"), we describe Andes2, our evaluations of it and theory-based explanations for the positive results. In the second subsection ("Research problems"), we describe the 4 main problems (Content, Scaffolding, Customization and Usability) that need to be solved in order to make Andes successful in a variety of high school classes. This sets the stage for the Project Narrative section, which indicates how we will achieve those goals.

1.1. Current State

1.1.1. The appearance and behavior of Andes2

The Andes2 user interface is intended to behave like pencil and paper. A typical physics problem and a partially completed solution 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 a physics problem with pencil and paper.

However, Andes2 goes beyond paper by offering the following features. As soon as an action is completed, Andes2 gives immediate feedback. Entries are colored green if they are correct and red if they are incorrect. This is often called flag feedback (Anderson, Corbett, Koedinger, & Pelletier, 1995). In Figure 1, all the entries are green except for the equation, which is red.

Unlike paper, variables must be defined. Paper does not require this kind of precision, so students often use variables in equations without defining them (and perhaps without understanding them). If students include an undefined variable in an Andes2 equation, the equation turns red and a message box pops up indicating which variable(s)

are undefined. Vector variables are defined in Andes2 by clicking on the tool bar on the left edge of Figure 1, dragging out an arrow, then filling out a dialogue box. Scalar variables are defined by filling out a dialogue box. The dialogue boxes require students to precisely define the semantics of the variables.

Andes2 includes a mathematics package. When students click on the button labeled "x=?" Andes2 asks them what variable they want to solve for, and then it tries to solve the system of equations that the student has entered. If it succeeds, it enters an equation of the form <sought-variable> = <value>.

Andes2 provides three more kinds of scaffolding:

- Andes2 pops up an error message whenever the error is likely to be a slip, which is defined as an error due to a lack of attention rather than a lack of knowledge (Norman, 1981). Leaving a blank entry in a dialogue box is an example of a slip. When an error is not recognized as a slip, Andes2 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*.
- 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*.

Thus, for errors that are likely to be careless mistakes, Andes2 gives unsolicited help, while for errors where some learning is possible, Andes2 gives help only when asked. This policy is intended to increase the chance that students will learn by independently repairing substantive errors without asking for help (eg. D. C. Merrill, Reiser, Ranney, & Trafton, 1992).

What's Wrong Help and *Next Step Help* usually generate a hint sequence. Most hint sequences have three hints. As an illustration, suppose a student who is solving a ramp problem has asked for *What's Wrong Help* on the incorrect equation *Fw_x = -Fs**cos*(20 deg)*. These are the three hints that Andes2 gives:

- Check your trigonometry.
- 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).
- Replace cos(20 deg) with sin(20 deg).

After each of the first two hints, Andes2 displays two buttons labeled "Explain more" and "OK." If the student presses on "Explain more," they get the next hint in the sequence. If the "OK" button is pressed, the problem-solving windows become active again, the lower left window becomes gray, and the student resumes work on the problem. This three-hint sequence is typical of many hint sequences. It is composed of a pointing hint, a teaching hint and a bottom-out hint:

- The *pointing* hint, "Check your trigonometry," directs the students' attention to the location of the error. If the student knows the appropriate knowledge and the mistake is due to carelessness, then the student should be able to pinpoint and correct the error given such a hint (Hume, Michael, Rovick, & Evens, 1996; D. C. Merrill et al., 1992).

- The *teaching* hint, "If you are trying to calculate…," states the relevant piece of knowledge in order to encourage just-in-time-learning. We try to keep these hints as short as possible, because students tend not to read long hints (Anderson et al., 1995; Nicaud, Bouhineau, Varlet, & Nguyen-Xuan, 1999). In other work, we have replaced the teaching hint text with either multimedia (Albacete & VanLehn, 2000a, 2000b) or natural language dialogues (Rose, Roque, Bhembe, & VanLehn, 2002). These more elaborate teaching hints significantly increased learning, albeit only in laboratory settings.
- The *bottom-out* hint, "Replace cos(20 deg) with sin(20 deg)," tells the student exactly what to do. It acts like a just-in-time-example. According to the model-scaffold-fade theory of skill acquisition (Collins, Brown, & Newman, 1989), this is a form of reverse fading (Atkinson, Renkl, & Merrill, 2003; Renkl & Atkinson, 2003). That is, when the scaffolding (pointing and teaching hints) fails, students may still learn from the model (the bottom-out hint).

As the student solves a problem, Andes2 computes and displays a score. The score is based mostly on the number of errors (red entries), the number of bottom-out hints and the various measures of the clarity of the solution. Andes2 puts little weight on the final answer because it provides such good help that students almost always get the right answer.

Andes2 can be used both offline and online. When used offline, students print their homework and hand it in on paper. When Andes2 is used online, students submit their problem solutions via the OLI learning management system (http://www.cmu.edu/oli). The Andes2 scores are sent to the instructor's grade book, which looks and acts like a spreadsheet. The cells contain the student's score on a problem as computed by Andes2; clicking on the cell displays the student's solution. The grade book can be dumped to a tab-delimited file that can be read by commercial spreadsheets and databases.

1.1.2. Evaluations of Andes2

Andes2 was evaluated in USNA every fall semester from 1999 to 2003. Andes2 was used as part of the normal USNA introductory physics course. The course had multiple sections. Each 25-student section was taught by a different instructor. Students in all sections took the same final exam and used the same textbook. However, different instructors assigned different homework problems and gave different midterm exams approximately monthly.

Experimental conditions: In sections taught by the 3 collaborating instructors, students were encouraged, but not required, to do their homework on Andes2.

Control conditions: Each year, the Andes2 instructors recruited some of their colleagues' sections as Controls. Students in the Control sections did the same midterm exams as students in the Andes2 section. Control sections did homework problems that were similar but not identical to the ones solved by Andes2 students. The Control instructors required students to hand in their homework, and credit was given based on effort displayed. These practices encouraged Control students to both do the assignments carefully and to study the solutions that the instructor handed out.

Learning measures: 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 three hours. The midterm 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 midterm exam questions (15 minutes each). On the final exam, students just entered the answer, while on the midterm exams, students showed all their work to derive an answer.

Midterm exam results: Figure 2 shows the midterm exam results (normalized via z-scores) for years 2000 through 2003. In all years, the Andes2 students scored reliably higher than the Control students. When pooled over all 4 years, the effect size was 0.61[1] overall and 1.2 on the most conceptual component of the scoring rubric. When the student's grade point average (GPA) is used as a measure of in-coming competence, there was no aptitude-treatment interaction. That is, low GPA and high GPA students both benefited from Andes2, and by equal amounts compared to doing their homework on paper. When students were divided into Engineering Majors, Science Majors and Other Majors, no interaction was found; all groups benefited the same amount.

Final exam results: The final exam covered the whole course. Because Andes2 coverage steadily increased over the years, we report its impact on the last final exam, when it covered 70% of the homework problems in the course, and compare the Andes2 student's scores to the scores of all the other students who took the exam. As Table 1 indicates, the Andes2 students' mean score were higher than the mean score of the non-Andes2 students ($d = 0.25$; $p = 0.028$). When students were split by majors, the Other majors did learn more with Andes2 than with paper homework ($d = 0.5$; $p = 0.016$) but the Science and Engineering majors did equally well with both Andes2 and homework.

Comparison with two benchmarks: In order to understand the significance of these results, it is helpful to compare them to two benchmarks. For intelligent tutoring systems, there are only a few in-school, semester-long, controlled evaluations in the open literature, and the widely-cited Koedinger et al. (1997) study is arguably the benchmark study against which others should be compared. Koedinger et al. evaluated a combination of an intelligent tutoring system (i.e., PAT) and a novel curriculum (i.e., PUMP). The combination is now distributed by Carnegie Learning (www.carnegielearning.com) as the Cognitive Algebra I Tutor. The Koedinger et al. evaluation and the Andes2 evaluation had remarkably similar effect sizes: 1.2 and 0.7 on experimenter-designed measures and 0.3 on standardized measures.

As a second benchmark, Andes2 can be compared to conventional homework systems such as WebAssign (www.webassign.com), Mastering Physics (www.masteringphysics.com), and LON-CAPA (www.LON-CAPA.org). These have students solve a physics problem on paper, then enter their answer into the systems. The system tells them instantly whether the answer is correct, and it may give a hint if it is not. Although the traditional name for these systems is computer-aided instruction (CAI) or computer-based training (CBT), a better name would be *answer-based tutoring systems* because they give students immediate feedback and hints on the final answer.

In contrast, Andes2 monitors every step the student takes while solving a problem, and gives feedback and hints on each step. Although systems like it are traditionally called Intelligent Tutoring Systems, a better name would be *step-based tutoring systems* (VanLehn, 2006).

[1] All effect sizes are computed with Cohen's d = (post_test_mean(experimental) – post_test_mean(control)) / standard_deviation(control). In order to aggregate scores across the 4 years, students scores were first convereted to z-scores.

We have not compared Andes2 directly with answer-based tutoring systems, but there is indirect evidence for its superiority. When students complete their homework on paper, it is often infeasible to grade every problem of every student. Thus, it is not surprising that when answer-based tutoring systems are introduced and the students' score on each problem now count in their course grade, then students tend to do more homework and to learn more. Fortunately, several studies have compared answer-based physics homework to paper-and-pencil homework while grading every paper-based solution (Dufresne, Mestre, Hart, & Rath, 2002; Pascarella, 2002, 2004; Bonham, Deardorff, & Beichner, 2003). All found null results on both experimenter exams and standard exams, including concept inventories. Thus, we have the following chain of results:

A. ungraded paper-based homework < answer-based tutoring systems
B. answer-based tutoring systems = graded paper-based homework
C. graded paper-based homework < a step-based tutoring system (Andes2)

Result C was discussed early in this section. Result A is probably due to students solving more homework problems with the answer-based tutoring systems than with ungraded paper-base homework. Andes should enjoy this benefit as well. The combinations of results B and C suggest that Andes should be more effective than answer-based tutoring systems by about d=0.61 for open response tests and d=0.25 for standard exams. The next section considers explanations for this "result" and for the third result C.

Student Attitude Surveys: One mark of the success of a teaching technique or tool is student acceptance. Each year that Andes2 was used at the USNA, students completed a survey instrument at the end of the semester. Figure 3 shows results obtained during fall 2005 and spring 2006 semesters. During this time, Andes2 covered 90% of the curriculum during the fall and 75% of the curriculum during the spring. The remaining homework assignments were completed as pencil-and-paper homeworks. Between the semesters, student sections were randomly re-assigned: thus, most of the students using Andes during the spring semester had used some other homework system during the previous semester. Students preferred Andes2 to other homework methods at the end of the fall semester and favored Andes2 much more strongly after completing the spring semester. There are several possible explanations for this result:

1. The spring semester students had more experience with other homework methods and could more knowledgably rate Andes2.
2. Andes2 may work more effectively for new material: many of the USNA students already completed a physics course in high school and, for those students, a substantial part of the fall curriculum was review, while more of the spring semester material was new. One would expect that a step-based tutor would not be perceived as beneficial for a student who has already learned the material.

1.1.3. Theory-based explanations for the success of Andes2

Given that all students are solving homework problems and their solutions affect their course grade, why do Andes2 students learn more than students who solve problems on paper? Our explanation is based on a detailed cognitive model of how students learn physics (Jones & Fleischman, 2001; VanLehn, 1999; VanLehn et al., 1992; VanLehn &

van de Sande). The model, named Cascade, is based on expert-novice studies in physics (Chi et al., 1981; Larkin et al., 1980; Priest & Lindsay, 1992), self-explanation in physics (Chi et al., 1989; Chi & VanLehn, 1991; Hausmann & VanLehn, 2007), tutoring of physics (Chi, 1996; Chi et al., 2004; Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001; Siler & VanLehn, 2003; VanLehn et al., 2003; VanLehn et al., 2005) and studies of transfer between quantitative and conceptual physics (Ploetzner & Kurt VanLehn, 1997). Cascade has been extensively evaluated (VanLehn & Jones, 1993; VanLehn et al., 1992) and its implications for Andes can be simply described.

First, if students do not get immediate feedback and hints, then they often fail to apply critical knowledge. For instance, on a problem intended to exercise Newton's second law, the student may generate and submit an incorrect solution without ever noticing that the law needs to be applied, so they lose the opportunity to refine their understanding of this extremely important and complicated law. Even though students in the USNA control condition got their paper-based solutions handed back with marks, they probably did not try again to re-solve the problem and thus exercise their knowledge. That is, once a paper-based solution is submitted without the student having applied a critical piece of knowledge, the student may have forever lost that particular opportunity to learn. Anderson, Corbett, Koedinger and Pelletier (1995) have argued that these lost opportunities for learning comprise the major explanation for the effectiveness of their Cognitive Tutors, which also give immediate feedback and hints.

Second, physics students have two main methods for solving practice problems in natural settings where they have access to previously solved problems (called "examples" henceforth, even if they don't appear in the instructional materials as such). One method could be called is s*uperficial analogical problem solving*:

1. Students search for an example with the same superficial appearance as the one they are trying to solve. E.g., both have inclined planes.
2. They form a superficial analogy between the example and the problem. E.g., a car rolls down a 30º driveway vs. a block slides down a 55º incline, so car::block, driveway::plane and 30º::55º.
3. They copy the equations from the example to the problem, changing them minimally in accord with the analogical mapping they formed. E.g., sin(30º) becomes sin(55º).
4. They generalize from this process to form a problem schema. E.g., For any inclined plane problem, $a=m*9.8*\sin(\theta)$ where m is the mass of the object sliding down the plane, and θ is the angle of the plane.

The second method could be called principle-based problem solving:

1. Students search for a principle (e.g., Newton's second law) that both applies to given situation and seems relevant to the sought and given quantities.
2. They write the equation that follows from applying the principle.
3. If that equation does not suffice to solve the problem, they repeat the above two steps.
4. They generalize from this process to form a principle schema. E.g., When a problem has two time points and asks about the change of motion of an object between them, and the problem seems to give enough information to calculate the total mechanical energy at each time point, then apply

Conservation of Mechanical Energy by calculating the kinetic energy and potential energy of the object at each point.

Both methods yield general schemas, but the analogical method produces a *problem* schema whereas the principle-based method produces a *principle* schema. During high-school algebra classes, students often learn problem schemas, and surprisingly few are needed in order to successfully solve all the word problems in the curriculum (Mayer, 1981). Similarly, students can "master" the physics curriculum by learning relatively few problem schemas. This allows them to solve most problems successfully, but they are often completely unable to explain what principles they used. Thus, they fail miserably on conceptual tests, such as Chi's problem sorting task (Chi et al., 1989) or Dufresne's problem similarity task (Dufresne, Gerace, Hardiman, & Mestre, 1992). When they give a verbal protocol as they solve problems, they rarely refer to quantities such as "velocity" or "acceleration" but instead refer to symbols such as *vi* or *a* because those symbols are the main constituents in the problem schemas. Thus, instructors often complain that even some of their best students are merely "pushing symbols" instead of doing physics.

For students to avoid acquiring a symbol-pushing "understanding" of physics, Andes should both discourage them from doing superficial analogical problem solving and encourage them to do principle-based problem solving. Andes2 is only partially successful in this respect. On the one hand, it reduces analogical problem solving by making it difficult to access some examples—the student must close the current problem in order to open a problem that was solved earlier. Andes cannot completely block access to examples, as students can always print them. However, by making example access inconvenient, it probably reduces the frequency of the analogical method. On the other hand, Andes2 tries to increase principle-based problem solving by giving principle-based hints. Whenever the student asks for help, Andes first tries to get the student to identify the principle to be applied. Only if that fails does it give a "bottom out" hint that tells the student what equation to write. Unfortunately, students do not read the initial hints and click rapidly on the help button in order to get to the bottom out hint. Thus, Andes2 encourages principle-based problem solving but it cannot force students to use it.

This is a second explanation for why Andes2 is more effective than graded homework. When students are solving problems on paper, they too frequently use the analogical method and thus acquire symbol-pushing, problem schemas. When they are solving the same problems on Andes, they are more likely to use the principle-based method and thus acquire deep, principle schemas.

To summarize, we believe there are 2 reasons why Andes2 students learned more than students solving problems on paper for a grade:

- If students fail to apply a critical piece of knowledge, Andes2 gives immediate feedback and hints until they do. If the student is instead working on paper, they will probably submit a solution without that knowledge being applied and thus have forever lost that opportunity to learn.
- Andes2 increased the frequency of principle-based problem solving and decreased the frequency of superficial analogy-based problem solving.

1.2. The research problems

An established principle of educational technology is that the design must be developed in collaboration with the instructors who will be using it. Developing

educational software in a lab and then "tossing it over the wall" to classrooms is a known recipe for failure. Thus, Andes2 and its predecessor Andes1 were developed from day one in collaboration with three physics instructors at the USNA: Profs. Robert Shelby, Mary Wintersgill and Donald Treacy. These instructors were active consumers of the Physics Education Research literature, and they were early adopters of instructional methods and materials advocated by Mazur (1993), McDermott, van Heuvelen (1991) and others. They routinely monitored student performance with the widely used concept inventories (Hestenes, Wells, & Swackhamer, 1992) in addition to standard multiple-choice assessments and their own open-response exams. They felt there was considerable room for improvement in their students' learning, which led to our collaborative development of Andes.

However, the USNA is an unusually homogeneous instructional environment. As we tried to get others to use Andes2, instructors very often were enthusiastic about the "Andes2 idea," but ran into difficulties when it came to actually using it in the classroom. Base on these failures (and recommendations from the instructors at Watchung Hills) we have recognize that Andes2 needs to be extended in 5 ways. We list them in the order in which instructors mention them.

1.2.1. Usability

All the USNA were provided with identical Microsoft Windows computers, software, and network connections. Thus, the Andes2 user interface was developed as a Microsoft Windows application program. This choice has greatly impeded its adoption in other contexts. Many schools have locked-down lab computers, making Andes2 installation problematic. In addition, an increasing number of instructors and students use other operating systems, enough to prevent at-home use for most classes. Thus, Andes3 needs to run as a web application; this is the single greatest barrier to widespread adoption. In addition, Andes2 has a fairly complex user interface, and it takes instructors about half a class period to introduce students to it. Alternatively students can watch an introductory video for interface training. Without this training, users can get quite frustrated. Andes3 needs to have a simple familiar interface that allows a user to start solving problems immediately, with no prior user interface training.

1.2.2. Support

Many instructors want a commercial level of support with some assurance that the product won't disappear in a few years. Other instructors want something that is free (currently the case for Andes2). We need to accommodate both cases, while allowing continued development of the software and course content.

1.2.3. Content

Andes2 has over 500 problems covering almost all the USNA year-long introductory course. Most of its problems are quantitative, like the one shown in Figure 1. Some are graphical, such as drawing all the forces acting on a projectile or drawing a light ray as it crosses a water/air boundary. Some problems are multiple choice conceptual problems, such as the ones used on concept inventories. However, a typical textbook has several thousand such problems, and even then, instructors cannot always find a problem that they want to assign. Indeed, whenever a college or high school instructor has signed on

to use Andes, they always ask us to add certain problems that they like to assign. Andes3 needs to find a solution to this.

Regular Physics classes make extensive use of conceptual physics problem, which are problems that do not require writing and solving a system of equations. For technical reasons, Andes2 does not have enough conceptual problems. Andes2 can handle only conceptual problems whose solutions can be expressed by filling in boxes with numbers or words, or by drawing certain kinds of diagrams. Andes3 needs to understand more complex diagrams (e.g., free-hand sketches of time-motion curves). Andes3 also needs to allow the student to type in explanations in natural language, and most importantly, Andes3 needs to understand the student's explanations. Fortunately, we can draw on nearly a decade of experience in building natural language tutoring systems for physics (Jordan, Makatchev, Papuswamy, VanLehn, & Albacete, 2006; VanLehn et al., 2007, 2005, 2002; VanLehn, Jordan, & Litman, 2007; Katz, 2006; Katz, Connely, & Wilson, 2007; Litman et al., 2006; Jordan, Rose, & VanLehn, 2001; Rose, Di Eugenio, & Moore, 1999). We know where the minefields lie and how to get around them.

1.2.4. Customization

Even on straightforward activity such as solving a traditional quantitative problem (e.g., "A 5 kg block slides down a frictionless plane inclined at 30 degrees from the horizontal…"), instructors have different practices. One instructor might prefer a parsimonious solution with just the key equations. Another might prefer to draw all the vectors, draw triangles for the trigonometry, write every equation in its generic form first, etc. As the three USNA instructors began to specify the behavior of Andes2, they realized that preferred problem solving practices were different. In consultation with the literature, they agreed on a common set of problem-solving practices, such as always drawing vectors instead of defining a vector variable in text. They agreed to use these practices whenever they demonstrated problem solving in lecture, and to use these practices as part of the homework and midterm exam grading rubrics. Thus, Andes2 was developed in an unusually homogeneous context of instructor practices.

Although instructors will refuse to use Andes2 on the grounds of usability or lack of content, they are merely irritated by the lack of control over the preferences expressed in its hints and grading rubrics. Andes3 should nonetheless give them control over these preferences.

1.2.5. Scaffolding

In general, the level of scaffolding provided by educational software should match the competence of the student (Collins et al., 1989; Vygotsky, 1978). Andes2 has several types of scaffolding: (1) feedback on steps, (2) hints on steps, (3) examples of solved problems, (4) required definition of variables, and (5) optional algebraic equation solving. For high school students, particularly those in Regular Physics, more scaffolding is required.

In summary, Andes3 will be a significant change from Andes2. While maintaining the basic policies of Andes2, we propose to develop a new homework tutoring system that will differ in Usability, Content, Customizability and Scaffolding. Table 2 summarizes the most important differences. Rows 1 and 2 refer to usability: Andes 3

will be accessible to high school students because it will be platform independent and delivered via the Internet. Rows 3 and 4 refer to content: Andes3 will contain physics problems that correspond to general high school introductory courses and will adapt to a variety of instructional practices. In particular, Andes3 will have many more problems similar to the ones it has now, and it will contain conceptual problems that are currently beyond its capabilities, such as those requiring understanding of student explanations or free-hand drawings. The next row, Row 5, refers to scaffolding: Andes3 will incorporate methods that have worked in lab studies and seem likely to work in the real world of Andes usage. The final rows in the table refer to participants and instructor support.

2. Project Narrative

Our research plan is based on the reality of supporting a growing user community while periodically making significant improvements to the system. This means managing a continuous process of small incremental improvements while also designing, pilot testing, implementing and formally evaluating much larger improvements. In Sections 2.1 through 2.3, we describe 3 major improvements that we anticipate making. As per the RFA[2], each is evaluated with a small experiment that compares the improved version of Andes3 to a version without the improvement. The evaluations will be conducted at a rate of approximately one per semester. The subsections follow the chronological order of the changes and their evaluations. After these major changes to the system are described, we describe the nature of the smaller, more incremental changes that we anticipate and discuss why they are important. This lead to a discussion of the user community, which is the source of these improvements and is foundational to the success of this project. The narrative concludes with our plans for feasibility evaluations.

2.1. New user interface and new computational architecture

Improving *usability* is one of the 4 major extensions required of Andes. In particular, students and instructors should be able to start working on Andes3 with minimal user-interface training and minimal software installation. This section describes how that goal will be met.

We hypothesize that these changes will improve the usability of the system without affecting its learning gains. The proposed evaluation will measure both usability and learning gains even though a thorough test of our null hypothesis about learning gains requires more subjects than we will be able to muster.

2.1.1. Andes3 computational architecture

In practice, Andes2's computational architecture has been the greatest obstacle to its widespread adoption. Andes2 has 3 main modules: (1) the front end, which the student interacts with; (2) the help system, which has the artificial intelligence that drives the system and (3) the math package, which provides mathematical services to the help-system. Although the help system and the math package are platform-independent, the

[2] Pg. 47: "Under the Technology program, typical Goal Two projects consist of a series of small experiments to determine which strategies…optimize learning."

front end is not. The USNA uses Windows machines and a fixed set of other computational infrastructure. Because Andes2 was funded by Office of Naval Research to increase learning at the USNA, we constructed a Microsoft Windows specific front end.

But in a high school, equity considerations are paramount and the software must run on every student's home computer. Although rewriting Andes to be platform independent is neither easy nor cheap, we have no alternative.

The new front end will allow students to access Andes3 on any computer that has a web browser and JavaScript installed. This includes other operating systems and computer lab machines that are "locked down." Our plan for achieving this is to use "AJAX" techniques. AJAX is an acronym for Asynchronous JavaScript and XML. It is a development technique for creating interactive web applications. Unlike classic web pages, which must load in their entirety when content changes, AJAX allows web pages to be updated asynchronously by exchanging small amounts of data with the server behind the scenes.

2.1.2. Andes3 user interface

The Andes2 user interface is complex. We have tested many techniques for user interface training and for just-in-time user interface help. At best, it still takes about 30 minutes to train a student or instructor initially. We need to eliminate this barrier, as it causes many potential users to abandon Andes prematurely.

The Andes3 user interface will to be like a piece of paper or a generic vector graphics drawing program (like PowerPoint) in that students can easily place text and graphics anywhere on it. For example, when starting to solve a quantitative problem, students might see a screen similar to the one shown in Figure 1. On the left side of the workspace is a column of buttons for drawing tools. From the top, the buttons are, respectively, for writing text, writing an equation, drawing a coordinate axes, drawing an arrow, drawing a straight line, drawing a curve or circle, drawing a dot and asking for help. The same button can be used for drawing different objects in different problems. For instance, an arrow can represent a vector in one problem or a light ray in another problem. Coordinate axes are used for vector drawings in one problem and for time-motion graphs in different problems. This is intended to keep the user interface constant so that students do not have to learn about new drawing tools when they start in on a new type of problem.

There are no buttons for defining variables. Instead, the student enters text, as in Figure 5. Andes3 will have limited natural language processing capabilities, so students may use a variety of ways of expressing definitions. If Andes3 cannot understand the student's text, it will pop up the help dialogue box and ask for clarification.

Most drawn objects require labels. After an object is placed or dragged out on the screen, a text box asking for the label appears. For instance, in Figure 5, after the student drew the body (represented as a circle), the tutor popped up box asking the student for the name of the object.

Whenever a student draws an object, it turns red if it is incorrect and green if it is correct. (Color-blind students can alter the color choices in a preferences menu.) In Figure 5, all but one of the student actions are colored green, showing that Andes3 has accepted them as being part of a correct solution.

The student can click on the help button at any time. If the student has just submitted an incorrect entry, the tutor will assume that the student wants help on it. Otherwise, it will assume that the student wants advice on what to do next. Asking for help brings up the help dialog box where the student can interact with the tutor (Figure 7). The tutor will give the student help in the form of a sequence of hints or sometimes by asking questions.

As the student is working on the problem, Andes3 scores the responses based on accuracy, amount of help received, the student's choice of steps and their sequence. The score is displayed in the top right corner (Figure 5). When students click on the "drop down" symbol in the score box, a form appears that explains how the score was derived and what the student needs to do in order to improve it. Instructors determine the exact scoring rubric, and this is a major way that they can customize Andes' pedagogy. For instance, if they want the student to draw all forces before entering any equations for Newton's law, then they can assign points to that measure. We have found that students pay a great deal of attention to their scores, and often want to know what they can do to improve them.

2.1.3. Development and evaluation

The task of constructing the Andes3 user interface began in April 2007 and continues under separate funding (from the NSF's Pittsburgh Science of Learning Center). In addition, we expect work on integration of Andes3 into the WebAssign and LON-CAPA homework systems to commence early 2009. Thus, we anticipate initial beta testing in spring, 2009 and a functional system in place by August 2009.

Over the summer of 2009, extensive user interface testing and refinement will be conducted in collaboration with the human factors team at WebAssign. Our main objective dependent measure will be the time to completion for a set of simple problems that the participants already know how to solve on paper. Testing and refinement will be conducted in several cycles. For each cycle, the high school liaisons, Megowan-Romanowicz and Gershman, will define a sample of tasks. Participants, both instructors and students, will solve the tasks first on paper then on either Andes2 or Andes3. Verbal protocols will be recorded in synchrony with the computer screen or videos of the paper. Experimenters will take notes, and probe students after each problem about apparent difficulties. Those difficulties will be reported to the development team, which will modify either the user interface or the user interface help. This will continue until there seem to be no remaining issue with these tasks. The cycle will then repeat with a new set of tasks. The cycles will stop when we reach diminishing returns, hopefully because users are able to start solving problems on Andes3 with no more difficulty than they would have in solving them on paper.

The evaluation of learnability will be conducted in fall, 2009. It will compare the two fronts ends, Andes2 and Andes3, for physics learning gains. The study will use a two-condition, between subjects design.

Sample: Participants will be drawn from our target population of Regular, Honors and AP physics students in American high schools. Approximately 8 teachers from our 5 collaborating high schools will participate—we will only use teachers who teach at least two classes at the same level (e.g., 2 Honors classes).

Procedure: One of their classes will be randomly assigned to Andes2, and the other to Andes3. The intervention will occur as part of the students' regular coursework. Students will use their assigned system for homework and seatwork up to the first mid-term exam. From then on, all students will use Andes2 (because only the first few units of the course will be available on Andes3).

Dependent measures: Log data will be used to insure that the intervention is occurring as planned. Dependent measures for testing the hypothesis will be scores on exam questions and learning curves. We expect different instructors to use different exams, but to use the same exam for both of their classes. Thus, exam data will be relevant to school learning and will test our hypothesis, but will not allow comparison across instructors.

We will encourage instructors to use open-response exams because we believe they are more valid than multiple choice exams for the kind of conceptual physics understanding that Andes tries to teach. During the USNA evaluations (VanLehn et al., 2005), we found evidence that the multiple-choice exams normally used for cross-class and cross-year assessments allow students to use either the methodical, principle-based problem solving method taught by Andes or the error-prone, symbol-pushing, superficial analogical problem solving method that many students tend to use (these two methods are described in more detail in Section 1.1.3). On the other hand, all students open-response exams tend to use the principle-based method that Andes teaches, thus making them a more valid assessment of it. That is why we observed different patterns of results from open-response and multiple choice exams in the USNA data.

Learning curves will be obtained by storing log data from Andes2 and Andes3 in the PSLC DataShop (VanLehn et al., 2007), which has tools for automatically computing learning curves. For each knowledge component (i.e., a fine-grained instructional objective, such as drawing a static friction force correctly or applying Newton's third law correctly), the learning curve generator locates all occasions in the log data of a student where that knowledge component could be used. It computes an assistance score for each occasion, which is a function of the number of errors and hint requests in the log before the student entered the knowledge component correctly. When aggregated across students, one often finds descending curves. The slope of the descent is a measure of the rate at which the students are learning. The faster their errors and hint requests decrease, the faster the students are learning the knowledge component. Learning curves will be used in this study and several others to be described later.

Data analysis: The data analysis will treat each instructor's pair of classes as a distinct experiment. Our target high school classes tend to be small (~20 students per class, so ~40 per experiment) and the variance on instructor exams can be high, so these are low-powered experiments. Each analysis will use the students' math/science GPA prior to entering the physics class (assuming it is available to us) as a co-variate in an ANCOVA.

We cannot meaningfully aggregate results across instructors/experiments, as different teachers may teach different levels of physics (Regular, Honors & AP) and may use different curricula, texts, exams, etc. Because our hypothesis is that there is no difference in learning gains between Andes2 and Andes3, will consider the hypothesis supported if all 8 experiments yield a null result. Although the logical of multiple comparisons might suggest using a Bonferroni correction, the class pairs could be

different enough tests that a negative result at $\alpha=0.05$ (i.e., Andes3 students had lower scores than Andes2 students) from any one of them should taken seriously, and it possible causes explored. We do not want to plunge blindly ahead with a user interface that could harm some students under some conditions. For similar reasons, we will also look for ATIs using a median split on incoming GPAs.

2.2. Goal scaffolding and partially solved problems

In general, the level of scaffolding provided by educational software should match the competence of the student (Collins et al., 1989; Vygotsky, 1978). As Andes3 is extended beyond the USNA, we anticipate increasing heterogeneity of students' competence. In particular, we expect less competent students; so stronger scaffolding will be needed. Improved scaffolding was one of the 4 main extensions to Andes discussed earlier.

2.2.1. Goal scaffolding

The first type of scaffolding is goal scaffolding. One can consider the solution of a problem to be a tree of goals, subgoals, subsubgoals, etc. The steps are the leaves in this tree. In several studies, students more rapidly acquired a cognitive skill if the goal tree was presented to them as blanks to be filled in (Singley, 1990). This was demonstrated in physics by Reif and Scott (1999), who developed a fill-in-the-blank tutor for Newton's law problems. Their tutor was step-based, but the students were given blanks to fill in with steps, and goal (step content) for each blank was explicitly stated. For instance, when students should write an equation for Newton's law, F=m*a, they are shown a template for an equation with 3 blanks labeled "sum of all force components," "mass," and "acceleration component." The Cognitive Tutors of Carnegie Learning make frequent use of such goal scaffolding in their high school math tutors. However, the only *labeled* blanks in Andes2 are the ones for the final answers. There are blanks for equations, drawings and variable definitions, but they have no labels. Students must figure out goals for themselves. Some students find this difficult. Andes3 will serve their needs better if goal scaffolding is available.

In Andes3, goal scaffolding will appear as gray text. Figure 7 shows a problem with goal scaffolding. This is what a student would see when first starting to work on the problem. The gray text indicates what to do but not how to do it. The gray text is linked to the solution graphs (VanLehn et al., 2005) stored in Andes3, so that when the student has achieved a goal, the corresponding gray text is removed from the screen.

2.2.2. Partially completed problems

A related type of scaffolding is a faded example. An example is just a problem with its solution, and a faded example displays part of its solution but lacks the rest. A faded example it is just a partially completed problem. Example fading has been shown to be effective in mathematics instruction (Schwonke et al., 2007; A. Renkl & Atkinson, 2003), and cognitive simulations suggest that it would be effective in physics as well (Jones & Fleischman, 2001). Andes2 allows users to store partially completed problems, and Andes3 will continue to allow that.

Although it should not be technically difficult to allows authors to create goal scaffolded and partially completed problems, having Andes create them would be more

convenient for instructors and easier for developers to maintain. Ideally, an instructor would first pick a problem, and then pick the level of scaffolding desired by scrolling through a sequence generated by Andes3. To the instructor, it should look and act like scrolling rapidly thorough PowerPoint slides. The first "slide" would be the completely unsolved problem. With each succeeding slide, more scaffolding is added, including both goal scaffolding and fully executed steps. The final slide in the sequence would a completely solved problem—an example. When the instructor finds an appropriate level of scaffolding for her students, she clicks a button to select it.

2.2.3. Development schedule and evaluation

Although there are some software and HCI issues to resolve, the major challenge is designing sequences of scaffolded problems that cause learning. Atkinson and Renkl (op. cit.) have compared forward fading to backward fading, and found that backward fading is slightly superior. However, they used sequences of isomorphic, 4-step probability tasks. It is not clear how to generalize their result to sequences of non-isomorphic, more complex physics problems. In trying to solve the puzzle, we will be guided by the basic Vygoskyan principle of keeping students in their zone of proximal development and by detailed cognitive modeling (e.g., Jones & Fleishman, 2001).

More importantly, we will be guided by extensive pilot testing. We expect the software to be ready in early spring, 2010. Focusing mostly on Regular physics classes, and using student volunteers in lab settings, we will explore various sequences of scaffolded problems during pilot testing. For instance, we hope to be able to replace, say, 4 problems that take 15 minutes each with a sequence of 20 scaffolded problems that take 2 minutes each, followed by the original 4 problems, now solved at 5 minutes each. Both the baseline and the new sequences take an hour, but the students should learn much more from the new sequence, and probably experience less frustration as well. We estimate that it will take most of the spring and part of the summer to understand this sequencing problem and develop both general prescriptions and specific instantiations.

In order to test our hypotheses, we plan to conduct a formal evaluation of a set of these scaffolding sequences during fall, 2010. This study has the same subject, procedure and data sources as the one just described (Section 2.2.1). That is, we will again use about 8 teachers, each teaching at least 2 classes at the same level so that we can use a between-subjects design with intact classes. The intervention will be conducted as part of the students' normal coursework. One class will use the new problem sequence, while the other class uses the old one. Units for the intervention will be selected so that they are bracketed by midterm exams, which will be used as pre and post-tests. These data will be supplemented by learning curves extracted from the Andes log data. We will also measure training time using log data. Our hypothesis is that the scaffolding-based sequences will improve learning gains without extending training time.

As in the first formative evaluation (Section 2.1.3), the class pairs are so different that aggregation or hierarchical linear models make little sense. Thus, we will again treat the 8 pairs as 8 independent experiments. Since they are all testing the same hypothesis, which we hope will be positive, the Bonferroni correction will be applied. Given the conservatism of the Bonferroni and the low power of each experiment, we will count the hypothesis as supported if even one of the 8 experiments has a non-null result.

2.3. Novel conceptual problem types

As mentioned earlier, extending the content of Andes3 is an important part of making it appropriate for high school students. Although a great deal of content will be developed within the existing Andes3 framework (see Section 2.4.1), there is an important class of conceptual problems that require considerable software development and a formative evaluation.

Although many conceptual problems have simple, multiple choice answers, they require moderately complex reasoning in order to derive those answers. For example, consider this problem:

> A dive bomber can release its bomb when diving, climbing or flying horizontally. If it is flying at the same height and speed in each case, and we ignore air friction, in which case does the bomb have the most speed when it hits the ground? (A) Diving. (B) Climbing. (C) Flying horizontally. (D) It does not matter. The bomb's impact speed is the same in all three cases. (E) More information is needed in order to answer.

Most novices choose A as their answer. The experts choose D. Their reasoning is:

1. The initial speed of the bomb in all 3 cases is the same, so the initial kinetic energy is the same.
2. The initial height of the bomb is the same in all 3 cases, so the initial potential energy is the same.
3. The initial kinetic and potential energies are the same in all 3 cases, so the initial mechanical energy is the same.
4. The only force acting on the bomb is gravitational force, which is conservative, so the final mechanical energy is the same in all 3 cases.
5. The final height of the bomb is the same in all 3 cases, so the final potential energy is the same in all 3 cases.
6. Thus, the final kinetic energy of the bomb is the same in all 3 cases.
7. And so the final velocity is the same in all 3 cases.

Each of these steps requires a qualitative application (not quantitative) of one of the quantitative relationship in the Conservation of Mechanical Energy schema (Larkin, 1983). Each such qualitative application is easy, and once they have all been made, the student can generalize and enrich the schema.

Unfortunately, Andes2 and most existing physics software can do little to support this reasoning besides posing the question, getting the student's answer (A, B, C, D or E), and giving feedback on it. Hunt and Minstrell's (1994) system, Diagnoser, goes one step beyond this. After eliciting an answer from the student, and before giving feedback, it also gives the students a short list of explanations. Each explanation is short—a sentence of two. This invites students to think a bit deeper about their answer than they would otherwise. However, we would like to do even more to elicit deep, qualitative reasoning from students.

2.3.1. Multi-step conceptual problem solving

Our approach will be to have students type in lines of reasoning. For instance, an ideal solution to the dive-bomber problem would have 7 textboxes, each containing some natural language similar to the language above. The dive bomber problem has a particularly long line of reasoning. Others problems often have only a few lines. The

key point is that each line is an application of a single relationship, e.g., the definition of kinetic energy, the definition of potential energy near earth, etc.

A certain amount of natural language understanding technology will be needed for Andes3 to be able to check the student's text against the text that it expects. We have faced a similar problem with our natural language dialogue systems for physics tutoring, Andes-Atlas (Rose et al., 2001; Siler, Rose, Frost, VanLehn, & Koehler, 2002), Why2-Atlas (Jordan et al., 2001; Rose et al., 2002; VanLehn et al., 2002, 2007), ITSPOKE (Litman et al., 2006), TuTalk (VanLehn et al., 2007), and Cordillera (Jordan et al., 2006). We have found some fairly simple techniques that work reliably on the kind of text expected, where words and word order both contribute heavily to meaning (i.e., the usual bag-of-word techniques, such as Latent Semantic Analysis, work poorly because "A is greater than B" is the same bag of words as "B is greater than A", and such comparative occur often in conceptual physics reasoning.).

It is critical that goal scaffolding and example fading be implemented before these kinds of conceptual reasoning problems. Students will definitely need to see faded examples and have goal scaffolding in order to learn how to solve this unfamiliar type of problem.

2.3.2. Development schedule and evaluation

The main technology development will take in the second year, leading up to beta testing in early spring 2011. This will leave the rest of spring and summer 2011 for pilot testing and refinement.

A formative evaluation similar to the others discussed earlier (Sections 2.1.3 & 2.2.3) will be conducted in Fall 2011. It will compare our conceptual problems to Diagnoser-like problems. That is, students in both conditions will answer the same conceptual problems. The treatment students will enter lines of reasoning on the Andes3 user interface. The comparison students will select answer and explanations from menus.

We will need to work especially closely with the collaborating instructors in designing this evaluation. Conceptual problems are often not used as homework or seatwork, but instead are use to provoke discussion in class. For provoking discussion, it probably doesn't matter how students enter the solution, so if Andes3 is only used for this purpose, it is likely to be no more effective than Diagnoser. On the other hand, given the kinds of scaffolding, feedback and hints that Andes provides, instructors may be willing to assign more conceptual problems for solution outside of discussion—they can only discuss a few problems in class, and this gives student a chance to think about many other conceptual problems. If instructors are willing to increase the number of conceptual problems assigned as homework or seatwork, then we'd expect Andes3 to be more effective.

2.4. Many small improvements

In addition to the three major changes to Andes, which were discussed in the preceding sections, we need to continuously make small improvements at the behest of the user community. The major types of improvements are described below.

2.4.1. New content

As mentioned earlier, extending Andes to a variety of high school classes requires adding many new problems. Although Andes2 has over 500 problems, a commercial textbook has at least three times that number. Our process will be simply to encourage instructors to send us requests for new problems, which the development team will then implement. Few high school teachers have the time required to author Andes problems themselves, but we would be happy to help any who do. The tools for authoring problems are good now and getting better—graduate student Sung-Young Jung is developing sophisticated debugging and knowledge engineering tools as part of his PhD work.

Andes is not intended to provide every problem type that instructors want to assign. Instead, we are making it available via WebAssign and LON-CAPA (see Section 2.5.2). Instructors should assign Andes problems only when they are arguably more effective than more traditional online problems, and we should avoid implementing Andes versions of existing problem types. For instance, until we implement the multi-step conceptual problems described earlier, Andes has no particular advantage over Diagnoser and other systems that merely give feedback on multiple-choice conceptual problems.

Because Andes3 has a highly graphical user interface, we should be able to implement multi-step graphical problems easily. For instance, several instructors have asked that Andes support sketching of time-motion diagrams and energy diagrams (Van Heuvelen & Zou, 2001). Unlike filling out worksheets with such diagrams, working on Andes allows students to get hints, feedback, goal scaffolding, etc.

In general, our philosophy is to rely on instructors for suggesting for new problems, including novel or unusual ones. They are more in touch with the needs of their students, so there is a good chance that the problems will be effective. Many instructors routinely invent new problems for their students anyway and some seem to enjoy it. Instructors are more likely to work with developers to refine and improve a problem or set of problems that they have invented themselves. Lastly, an instructor who has authored several problems for Andes, and perhaps even seen their name on the web as the problems' author, will probably be a more active participant in the Andes3 community.

2.4.2. Customization

We have found that different instructors have different preferences for acceptable steps leading up to an answer. At the USNA, for instance, instructors want students to define all variables explicitly, so that is what Andes2 requires. Some instructors would prefer to drop this requirement, which would mean that Andes3 must be able to infer the semantics of a variable from its name and its use in the equations (Liew, Shapiro, & Smith, 2004). Here are some more examples of differing instructor practices:

- Some instructors want students to introduce the main steps in solving a problem with a phrase such as "Now we apply Newton's second law." Andes2 allows this but does not require it.
- Some instructors want students to solve the equations algebraically themselves; others are happy to let Andes3 do that work for the students.
- Some instructors want students to write vector equations for principles even though the equations cannot be solved—they must be first converted to scalar equations involving components of the vectors.

- Before submitting a solution that includes vector diagrams, some instructors want students to adjust the lengths of the vectors to more accurately represent their magnitudes.

For Andes to honor instructor preferences, we intend to use a large and expanding set of scoring rubrics. Instructors can award points for the solution conventions they want students to obey. If they cannot find a rubric that represents their pedagogical policies, then they can write their own rubric (on paper) using the existing ones as models. This should give the developers a head start in implementing the rubric.

2.4.3. Development and evaluation

We have been making many small changes to Andes for years, so management processes are already in place. Briefly, all feature requests are entered in Bugzilla; interesting ones are put on the agenda and discussed at the next Andes3 meeting; developers build private versions of the system for testing by the feature's requester; new releases of Andes occur every 2 or 3 months. It is infeasible to formally evaluation these changes, of course, but we at least try to get multiple instructors to try out and endorse the non-trivial ones before releasing them.

Sometimes it seems that the key to good educational software is not the overall design or the theory behind it, but the number of small, user-driven improvements that have been made. If so, then the key to success is to grow and support an active user community who will suggest and test such improvements. That is the topic of the next section.

2.5. Building and supporting the Andes3 community

For the development of Andes2, we relied on a committee of instructors, researchers and developers that met bi-weekly for a teleconference to discuss problems, design experiments, and coordinate research efforts. We plan to continue this practice for Andes3. However, with an increased number people involved in the project and multiple time zones involved, we need to find additional methods of communication. This mechanism is intended to support the core team, although we will encourage all our instructors and developers to attend if they have the time and interest. The rest of this section presents other elements of our plan for growing and supporting the community of users (first subsection) and the community of software developers (second subsection).

2.5.1. Community of Users

In order to communicate with the broader user community and encourage new teachers to join the effort, we will sponsor semi-annual workshops, to be held in conjunction with American Association of Physics Teachers (AAPT) national meetings. At these meetings, experienced Andes users will test the latest version of Andes3 and offer us feedback. Also, the workshops will offer training to get new users started, introducing teachers to Andes3 and offering them advice for fitting Andes3 into their curriculum and teaching goals.

Each year we will host instructors for a week-long workshop at ASU. During this time, instructors and researchers work in close collaboration, to develop new course content and problem types for Andes3. (Qualitative analysis of the feasibility evaluation data, discussed later, will also occur during these workshops.) Since Andes3 is supposed

to be used by a wide variety of classes, we plan for 9 (almost all) collaborating instructors to be involved with this. This involves analyzing the existing Andes curriculum, formulating new problem definitions (usually custom made, for copyright reasons), and pilot testing the resulting Andes problems

Finally, we need to provide a method of asynchronous communication where teachers can communicate with one another and with the research team. We will use a web-based forum, moderated by Prof. Megowan-Romanowisc, with an associated mailing list, to facilitate communication between users of Andes3 and members of the project team. As Andes3 matures, we expect this to become the primary means of helping new teachers getting started with Andes3.

To determine the effectiveness of the Community of Users, we will measure, growth, participation, and the effectiveness of the help-giving process. Our growth target is to have 20 teachers using Andes3 extensively by the end of the project. More importantly, we need to determine the growth mechanism: is it exponential or merely linear? We will survey new Andes users annually to determine how the instructor heard about Andes3, what caused them to try it, the support given and suggest ways of improvement. For participation, we will record attendance at the teleconferences (target: 5-10 people), experienced user participation in the workshops (4 for year 1, 8 for year 2, and 16 for year 3). We expect that participation in the web forum will remain constant or increase slowly (Lakhani & Hippel, 2003). To measure help-giving effectiveness, we will look at the time between when a query is first posted to the bulletin board or E-mail list, and the time that a substantive response is given.

2.5.2. Community of Developers

The software behind the Andes system is large and complicated; it is the result of over 20 person-years of work, with substantial contributions from over a half dozen people. In order to maintain the long-term viability of the software, it is essential that we foster a community of developers who understand the code, can maintain it, fix bugs, and make further improvements. To this end, we have entered into partnerships with LON-CAPA and WebAssign to cooperate on software development. WebAssign is the leading commercial online physics homework system (LMS) while LON-CAPA is the leading non-commercial equivalent. To facilitate cooperation, the Andes system software will be made available to the general public under an Open Source license. We feel this will be the best way to ensure that all interested parties can freely contribute to maintenance and development of the software. In order to foster the next generation of developers, we propose to train graduate students in physics and in Educational Technology to understand and further develop the software. Since this is a technologically sophisticated group, communication will occur mainly via E-mail and collaborative software development tools, such as Bugzilla (www.bugzilla.org). Finally, to foster community among the developers and coordinate the work among our partners, we will sponsor an annual developers meeting, to be held either in East Lansing MI or Raleigh NC.

Our software development plan follows an open source model. According to this model, a successful project will experience an exponential growth in the number of users as a function of time, while the total number of support requests must remain constant or rise linearly in time (Lakhani & Hippel, 2003). Thus, we propose two outcomes to measure the success of the software development process:

1. exponential growth in the number of users as a function of time, with
2. the total number of bug reports, requests for content, and requests for new functionality remaining roughly constant in time.

2.6. Feasibility testing

The three formal evaluations mentioned so far compare one version of Andes to another. That is helpful for deciding if a major feature is worth scaling up, and it helps inform educational theory and technology development. However, these evaluations leave unanswered an important, common sense question: Is Andes more effective than the paper and pencil (or whatever) technology that it replaces? If so, does its value-added exceed its adoption costs?

A complete answer to these questions would require a Goal 3 or 4 IES project. Nonetheless, it is important to at least check for gross violations of our expectations.

Our plan is to use cross-year comparisons to check for gross signs of infeasibility. When we have located an instructor who is interested in possibly adopting Andes, we will arrange to study their current class, which is presumably doing their work on paper or a web-based homework grading service such as WebAssign. After obtaining the necessary IRB approvals and permissions from the students and their guardians, we would arrange to observe the classes, interview students and the instructor, collect photocopies of exams. We will also collect a complete set of assignments, including homework and in-class work, and determine which ones can be replaced by Andes tasks.

Over the next summer, we will augment Andes as necessary to support more of the instructor's assignments. We will spend at least a day with the instructor sometime during the summer, working through Andes problems. This day serves several purposes. It familiarizes the instructor with a range of Andes problems and allows them to revise their syllabus as they see opportunities to do so. With our help, they will customize Andes to reflect their pedagogical preferences (see Section 2.4.2). This includes adjusting the grading rubrics to encourage the students to use the desired problem-solving methods.

During the next school year, the instructor will use Andes3 for the first time. By default, we will provide on-site support during the first few days of usage, although some instructors may ask us not to bother, if they feel especially comfortable with Andes.

Throughout the year, we will collect the same data as we collected during the baseline year, plus log data from Andes. The same intensive data collection will also occur in the third year. Thus, for each instructor, we will have a corpus of data from a baseline year, and the first two years of Andes usage.

Analyses and interpretations of these data are made difficult by several factors. Physics classes tend to be small, about 20 students or less. Although the state mandated testing and the concept inventories are comparable across years, the open-response exams that we believe are more valid measures of principled-physics competence, tend to be different from one year to the next.

Because we need to compare a variety of qualitative data across years when both the students, exams and instruction are changing, we cannot draw strong inferences about whether Andes or other sources caused the observed differences. Nonetheless, something must be done in order to see patterns. Our plan is to ask the Andes3 community to write interpretations of each 3-year corpus. Prof. Megowan-Romanowisc has designed and

moderated similar community-based, qualitative interpretation workshops as part of her leadership of the "Modeling Physics" physics reform program (modeling.asu.edu).

A likely method is to divide the three instructors attending the one-week workshop mentioned earlier into separate conditions. Each instructor, working with a wiki that exposes data from the other instructors, develops a preliminary interpretation/redescription/story for each year of the other instructors. The 3 instructors then meet to discuss and merge their stories. The goal is to find a single narrative, probably embedded in the data wiki, that spans each of the 3 classes and the 3 years, with sub-narratives for each. The narrative from each 3-person group are then presented to the other groups of instructors, with the goal of developing an overall, consensus story of what happens when Andes is introduced.

We believe this can be accomplished during 3 days of the one-week workshop. During the first day, instructors will meet briefly in their three-person groups, then spend most of the day developing individual narratives. Day two will be spend in the 3-person groups merging these narratives. Day three will be spend meeting as a whole group, discussing how to formulate a single narrative.

3. Personnel

The research team combines expertise in intelligent tutoring system development and evaluation (VanLehn, van de Sande), expertise in implementing and evaluating new instructional technologies in high-school classrooms (Risley, Kortemeyer, Megowan-Romanowicz), expertise in teaching physics at the high school level (Megowan-Romanowicz, Gershman, and other high-school teachers), expertise in knowledge-rich software and AI (van de Sande, VanLehn, Jung), and expertise in web-based educational software development along with the associated usability testing (WebAssign, LON-CAPA, new hire). More specifically, the team consists of:

Leadership and roles:

- Prof. Kurt VanLehn (PI), Arizona State University: managerial support, develop methods for determining Andes3 customizations appropriate for each instructor.
- Dr. Brett van de Sande (co-PI), Arizona State University: implement code to parse student quantity definitions; add code that allows customization features to change hinting behavior; manage communication between the UI and the tutor system; promote Andes3 in the Physics Education Research Community

Other personnel and roles:

- Colleen Megowan-Rabowisc, Arizona State University: liaison with Arizona high schools; instructor community development; training instructors; evaluating new Andes3 course content; administering and analyzing student questionnaires.
- Prof. Gerd Kortemeyer, Michigan State University (consultant): Advice on integrating Andes3 into LON-CAPA.
- John Risley, consultant: liaison with NC High Schools; help conduct Andes3 workshops at AAPT conferences; lead integration of Andes3 in WebAssign.
- Sophie Gershman, consultant: organize and help with semi-annual AAPT workshops; workshop follow-up, including training new instructors; review and suggest new course content.

- Sung-Young Jung (graduate student researcher in computer science) develop tools for expediting acquisition, maintenance and debugging of the Andes3 physics knowledge base; develop natural language input capabilities for variable definitions and qualitative questions; conduct studies to determine the effectiveness of the natural language-related features.
- Graduate student researcher in physics education (to be hired): Develop new problem content, respond to content and help-related bug reports, develop new problem types, conduct experiments to testing hints and new problem types.
- Graduate student researcher in Educational Technology GSR (to be named): develop methods to allow instructor customization; analyze student log files (*Usage Studies*); conduct studies to determine efficacy of new features of Andes3.
- Software engineer (to be hired): The development of Andes3 will include design and implementation of a JavaScript based user interface (UI) that will be delivered over the Internet, using "Ajax" technologies. This will require us to hire a programmer who has expertise in this kind of web-based programming. Tasks include: develop the new UI (6 months); addition of new problem types (in particular, graph drawing) (12 months); implementing code that determines the correct Andes3 customizations for each instructor (6 months); help with integration into LMSs (6 months); help with usability testing (6 months). Quality control for these tasks will be provided by lab studies, instructor bug reports, and student log file analysis. The programmer will also respond to bug reports associated with the user interface.

4. Resources

Arizona State University will provide office space, cyber infrastructure, and lab space for in-house usability studies of Andes3.

Andes3 will be delivered to students over the Internet through our partners, WebAssign and LON-CAPA. In addition, our partners will contribute to the development of the software, and advise on usability issues.

Our partnering High Schools offer a high level of diversity:

1. Watchung Hills Regional High School, Watchung NJ, high-income public. Race/Ethnicity: 80% White, 13% Asian, 5% Hispanic.
2. Enloe High School, Raleigh NC, public magnet school. Race/Ethnicity: 47% White, 35% Black, 12% Asian.
3. Broughton High School, Raleigh NC, public magnet school. Race/Ethnicity: 58% White, 29% Black, 5% Hispanic.
4. North High School, Phoenix AZ, low-income public. Race/Ethnicity: 68% Hispanic, 16% White, 9% Black, 5% Native American.
5. Bellevue Christian School, Bellevue WA, high-income private. Ethnicity: mainly white.

All instructors have ties to physics education research groups at nearby universities (Rutgers U., U. North Carolina, Arizona State U., and U. of Washington). We have letters of support from principals at all five schools, but have included only 2 due to the space requirements.

Appendix A (15 pages)
Figures, Charts, Tables, Examples of measures, Letters of agreement

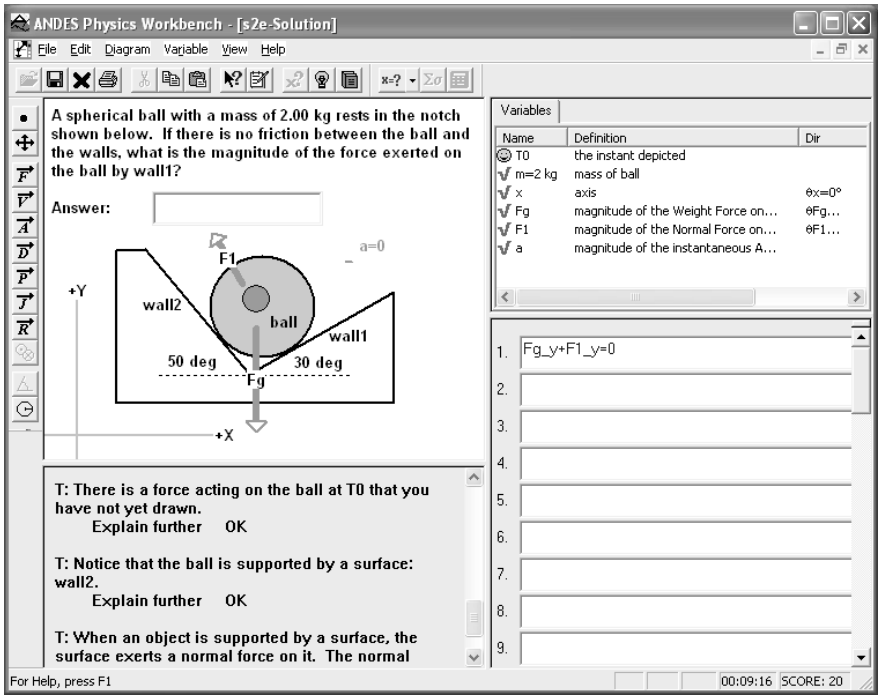

Figure 1: Screenshot from the current homework tutor, Andes2. The screen has areas for the problem statement and student-drawn diagrams, defining quantities, and entering equations. The lower left area is where the tutor gives hints. Student entries turn green if it is correct or red if incorrect. In this example, the equation has turned red and the student has subsequently asked for a number of hints.

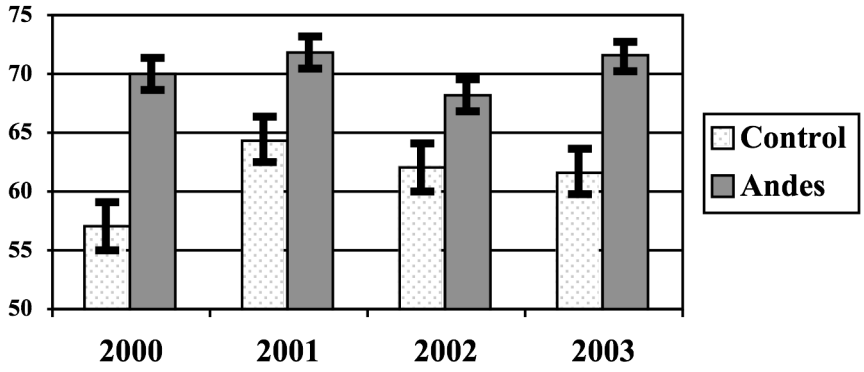

Figure 2: Midterm exam results from the USNA. The control condition students had (mostly) graded pencil and paper homework and the Andes condition used Andes2 for homework. All differences are reliable $p<.01$.

1. Was Andes more effective for doing homework versus doing homework in a more traditional fashion and turning it in?

	More				**Less**
	1	2	3	4	5
Fall 2005	24%	34%	11%	14%	15%
Spring 2006	48%	21%	17%	10%	5%

3. Do you think you learned more or less using Andes than if you had done the same exercises with pencil and paper?

	More(with Andes)				**Less**
	1	2	3	4	5
Fall 2005	18%	32%	7%	23%	1 8%
Spring 2006	37%	24%	17%	14%	8%

4. Would you have done the same number of exercises correctly with pencil and paper?

	More(with paper)				**Less**
	1	2	3	4	5
Fall 2005	17%	8%	28%	23%	24%
Spring 2006	3%	16%	16%	38%	27%

30. Compare your experience with Andes with your knowledge of other learning tools used at USNA like WebAssign, Blackboard or the on-line Chemistry assignments. Please circle the tool used for comparison.

Andes was	**More effective**				**Less effective**
	1	**2**	**3**	**4**	**5**
Fall 2005	31%	30%	15%	11%	13%
Spring 2006	44%	32%	16%	5%	3%

Figure 3: End-of-semester student questionnaire results collected at the USNA from students who had used Andes2. 71 students in Fall 2005 and 63 students in the Spring 2006. Questions relating Andes to other homework methods are shown.

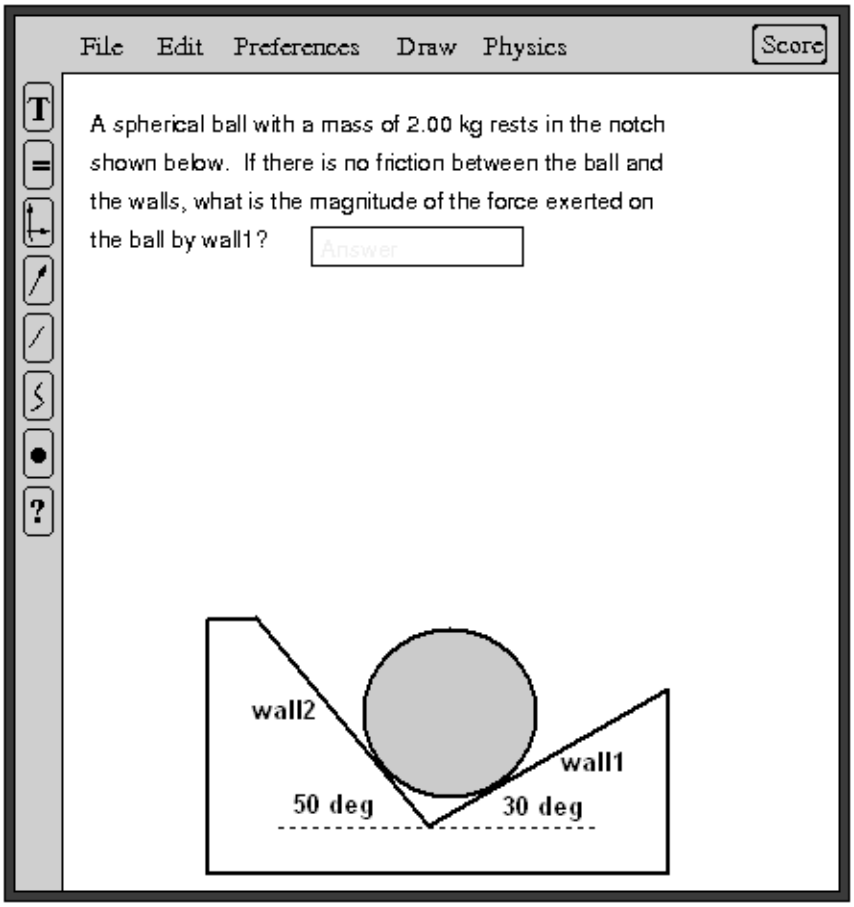

Figure 4: Artist's conception of Andes3 interface showing a problem statement before the student begins working.

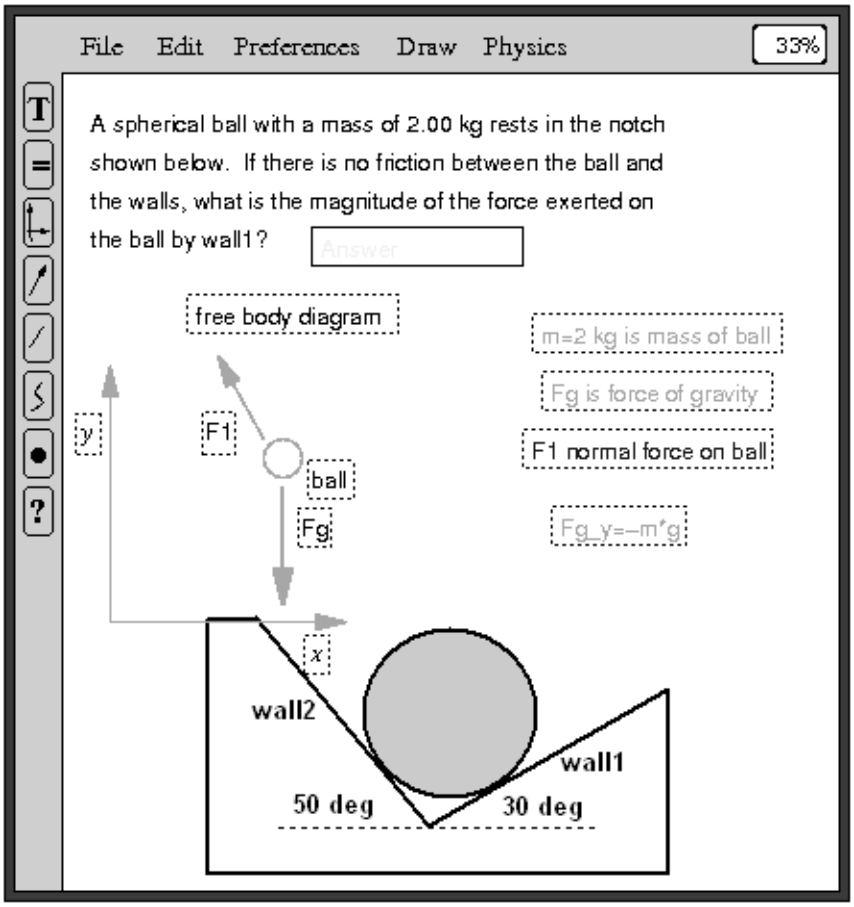

Figure 5: Artist's conception of Andes3 interface showing partially completed problem. The student entries – solution steps – turn green, showing correctness. In this example, the definition of F1 has turned red, indicating an error.

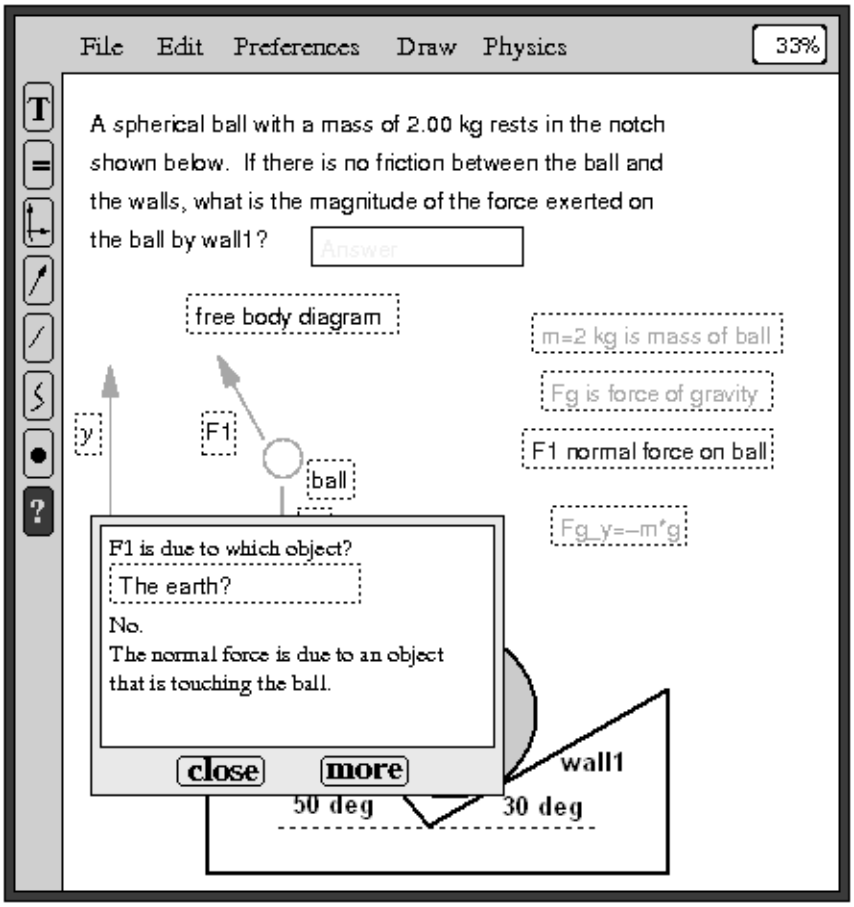

Figure 6: Artist's conception of Andes3 interface showing help dialog box. The help system has the capability for interactive dialog, although most hints will be declarative: the student clicks "more" for additional hints or "cancel" when they are done.

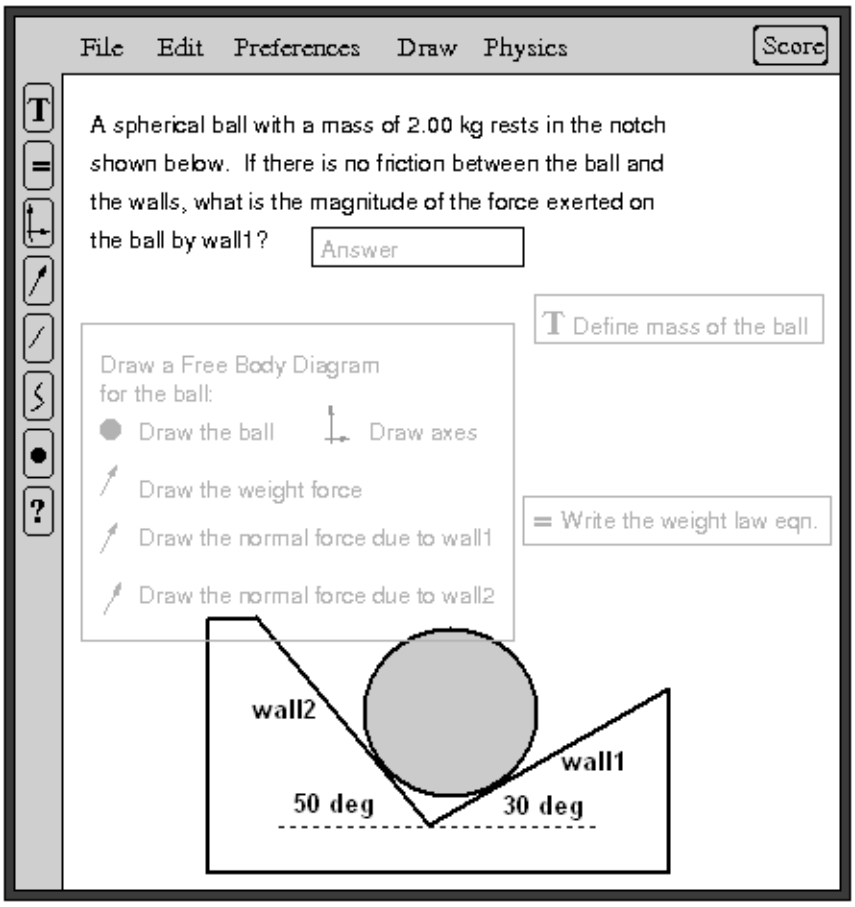

Figure 7: Artist's conception of Andes3 interface showing a goal-scaffolded problem. The gray instructions correspond to steps the student takes when solving a problem. The instructions may serve as a guide (for example, the directions of the force vectors are not given), rather than complete step-by-step instructions. As the student completes a step, the corresponding gray instruction disappears.

	Andes2	**Andes3**
Platform & delivery	Windows application	All platforms; client-server
User interface learnability	Required 30 minute training	Very little
Problems	About 500	Add: 1000+
Problem types	Quantitative, vector drawing, multiple choice	Add: natural language explanations, time-motion graphs, etc.
Customization	None	Instructors control grading rubrics and some hints
Scaffolding	Feedback on steps, hints on steps, examples, required variable definition, algebraic equation solving	Add: self-explanation prompting, many others
Participants	US Naval Academy, 1 high school	Add: Numerous high schools and colleges, community of users
Support structure	In-house	Multiple LMS providers

Table 1. Major differences between Andes2 and Andes3.

Residual scores on the 2003 final exam				
	Engineers	**Scientists**	**Others**	**All**
Number of Andes students	55	9	25	89
Number of non-Andes students	278	142	403	823
Andes students mean (stand. dev.)	0.74 (5.51)	1.03 (3.12)	2.91 (6.41)	1.38 (5.65)
Non-Andes students mean (s.d.)	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	0.520	0.25

Table 2. Andes2 students' mean score for the 2003 final exam was higher than the mean score of the non-Andes2 students. All types of majors benefited equally.

Response to prior reviewer feedback

The previously submitted proposal received thoughtful and well-deserved criticism from the reviewers. In response, we have made a number of significant changes to the proposal as summarized below.

From the Summary:

Discussion centered on the issue of feasibility. The panel felt that data would be available to study feasibility. However, there is no clear sense of how the data will be organized or analyzed. The plan stresses formative evaluation of software as opposed to feasibility for use in schools. From reviewer B: *The RFA identifies that an application should present methodologies to test the feasibility of implementation in an authentic setting and whether the intervention works as intended (implementation validity). ... information on these facets of the research plan is omitted. Description of methodology for analyzing data for pilot studies in the section "Non-exempt pilot studies research narrative" is nominal. Sample characteristics (e.g., prior knowledge, time in the semester/school year when pilot studies are carried out) and number of participants are not identified. It is not evident how some features of Andes3 (e.g., capability of feedback to promote learning) can be examined under the proposed procedures.*

The Project Narrative now includes explicit discussion of feasibility and related evaluations in Sections 2.6, 2.1.3, 2.2.3, 2.3.2 and the ends of 2.5.1 and 2.5 2.

The panel stressed that the RFA requires specificity about type of high school sample, yet the application was not clear in describing the sample. ... In addition, the applicant must specify the type of learner (advanced placement or other) to be targeted. From reviewer A: *The investigators do not specify their target audience—if they have interest in helping high school communities who traditionally struggle with science and physics, there should be more review of what is already been attempted with those groups. If their interest is in higher end kids, this should be spelled out.*

In several places, we have added a description of the courses and students that will be targeted during this study. Also, the demographics of the student population are described in the Resources section.

The panel urged that personnel must include someone who knows a good deal about high schools. From reviewer A: *None of the personnel appear to have clearly identified research expertise with struggling high school students in science.*

We have recruited Prof. Colleen Megowan-Romanowicz, who has over 20 years experience teaching physics in High Schools, as well as experience in introducing new technologies into the high classroom as project leader for David Hestenes' "Modeling Physics" project. In addition, we plan to expand the advisory role of Sophie Gershman, a high school teacher with over 20 years experience. Finally, we should mention that both Risley and Kortemeyer have experience in introducing technology into high-school physics classrooms: WebAssign and LON-CAPA are used by tens of thousands of high school physics students.

Additional comments from reviewer A:

Also, content for an AP-level group of kids and kids with more challenges is quite different—it is not made clear how the content to be built will be chosen: Which topics,

which pedagogies, which levels of mathematical expertise, or what the criteria for mastery will be. The focus appears to be on making a flexible tool that teachers can configure as they wish, but no discussion of whether some selectivity around methods and approaches is warranted.

Note that we are not proposing a curriculum; Andes3 provides computer-supported activities that instructors can use in a wide variety of curricula as replacements for paper and pencil activities (e.g., homework) or as supplementary activities. As explained in the Section 2.4.1, new content will be created following instructor requests: this method worked well for Andes2 and we see no reason to do otherwise for Andes3. Also, we are not (in this study) going to provide mastery-based learning. In general, the teacher will decide what problems to assign, what the grading rubric is and mastery criteria are, and when to advance students. With regard to selectivity of methods and approaches, we have *never* had a physics instructor ask us for help in selecting methods or approaches. The physics instructors we know all seem to have strong ideas about what they would like to do in their classes, and they are often avid readers of the Physics Education Research literature. Our ambition is merely to help them achieve their goals by providing a more effective media for problem solving than paper.

At the start of the application, there are few of the usual citations about the (well-known) problems of science and physics mastery in U.S. schools.

We have added some relevant citations.

The investigators do not indicate how they will ensure the writing and images will connect with high school learners with challenges. They do not specify how much new content will be built through this budget. … but needs to be more specific about what content and methods, exactly, are to be added that fit high school students with problems and more details on how the actual development will take place.

As detailed in Section 2.4.1, new content is suggested by the teachers. It is hard to estimate how many problems will be added without knowing what they want. If Andes can solve a new problem with its existing knowledge, then it takes only minutes to add the problem to Andes. On the other hand, if new knowledge must be formalized and debugged, adding the problem can take a day or even a week. To give one some idea of what we could do if needed, 300 problems were added to Andes2 over one year by one knowledge engineer (van de Sande). Most of these problems required extensive knowledge engineering, as they were from physics topics (e.g., electricity, magnetism, optics) that Andes had not previously covered. Now that Andes covers almost all of introductory physics, we hope that most of the new problems are the minute-long type rather than the day-long type.

At a high level, their description of what they plan to do for feedback seems good, but there are few details on methodologies to evaluate further; e.g., number of students, methods of coding or analysis. … they are not specific about how much new content will have to be evaluated or about the specific methods (e.g., numbers of students) to be used to get usability feedback for improvement.

The Project Narrative now includes those important methodological details. However, we should point out that most new content (e.g., new problems) falls under the category of small incremental changes, and thus their impact on learning will not be evaluated.

Unless the software engineer has the background, it is unclear who brings expertise in usability or interface design to the project.

The software engineer is expected to have background in this area. Also, our LMS partners WebAssign and LON-CAPA have considerable practical experience in interface design and usability. WebAssign has a staff member whose specialty is evaluation of user interfaces for usability.

Additional comments from reviewer B:

No CV is provided for the second Codirector (Dr. Gerd Kortemeyer).

We have provided CVs for all faculty-level personnel associated with the project.

The requested funds for programming seem somewhat underestimated.

We have designated additional funds "Integration into LMS" for incorporating Andes into WebAssign and LON-CAPA. Also, we have transferred some of the programming work to GRAs. Finally, we expect to share the programming work with our partners WebAssign and LON-CAPA.

Appendix B (10 pages):

Examples of materials for intervention and assessments

Bibliography

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

Albacete, P. L., & VanLehn, K. (2000b). The Conceptual Helper: An intelligent tutoring system for teaching fundamental physics concepts. In G. Gauthier, C. Frasson, & K. VanLehn (Eds.), *Intelligent Tutoring Systems: 5th International Conference, ITS 2000* (pp. 564-573). Berlin: Springer.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive Tutors: Lessons Learned. *The Journal of the Learning Sciences*, *4*(2), 167-207.

Atkinson, R. K., Renkl, A., & Merrill, M. M. (2003). Transitioning from studying examples to solving problems: Effects of self-explanation prompts and fading worked-out steps. *Journal of Educational Psychology*, *95*, 774-783.

Bassok, M., & Holyoak, K. J. (1995). Judging a book by its cover: Interpretative effects of content on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *23*(3), 354-367.

Bonham, S. W., Deardorff, D. L., & Beichner, R. J. (2003). Comparison of student performance using web and paper-based homework in college-level physics. *Journal of Research in Science Teaching*, *40*(10), 1050-1071.

Chi, M. T. H. (1996). Constructing self-explanations and scaffolded explanations in tutoring. *Applied Cognitive Psychology*, *10*, S33-S49.

Chi, M. T. H., Bassok, M., Lewis, M., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *15*, 145-182.

Chi, M. T. H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*(2), 121-152.

Chi, M. T. H., Roy, M., & Hausmann, R. G. M. (2008). Observing tutorial dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning. *Cognitive Science*, *32*(2), 301-341.

Chi, M. T. H., Siler, S., & Jeong, H. (2004). Can tutors monitor students' understanding accurately? *Cognition and Instruction*, *22*(3), 363-387.

Chi, M. T. H., Siler, S., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science*, *25*, 471-533.

Chi, M. T. H., & VanLehn, K. (1991). The content of physics self-explanations. *The Journal of the Learning Sciences*, *1*, 69-105.

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the craft of reading, writing and mathematics. In L. B. Resnick (Ed.), *Knowing, learning and instruction: Essays in honor of Robert Glaser* (pp. 453-494). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dufresne, R. J., Gerace, W. J., Hardiman, P. T., & Mestre, J. P. (1992). Constraining novices to perform expert-like problem analyses: Effects on schema acquisition. *The Journal of the Learning Sciences*, *2*(3), 307-331.

Dufresne, R. J., Mestre, J. P., Hart, D. M., & Rath, K. A. (2002). The effect of web-based homework on test performance in large enrollment introductory physics courses. *Journal of Computers in Mathematics and Science Teaching*, *21*(3), 229-251.

Elio, R., & Scharf, P. B. (1990). Modeling novice-to-expert shifts in problem-solving strategy and knowledge organization. *Cognitive Science*, *14*, 579-639.

Ewald, G., Hickman, J. B., Hickman, P., & Myers, F. (2005). Physics First: The Right-Side-Up Science Sequence. *The Physics Teacher*, *43*(5), 319-320.

Hausmann, R. G. M., & Chi, M. T. H. (2002). Can a computer interface support self-explaining? *Cognitive Technology*, *7*(1), 4-14.

Hausmann, R. G., & VanLehn, K. (2007). Explaining self-explaining: A contrast between content and generation. In *Proceedings of AI in Education, 2007*. Amsterdam: IOS Press.

Hausmann, R. G. M., van de Sande, B., van de Sande, C., & VanLehn, K. (2008). Productive Dialog During Collaborative Problem Solving. *0806.0599*. Retrieved June 4, 2008, from http://arxiv.org/abs/0806.0599.

Hehn, J., & Neuschatz, M. (2006). Physics for All? A Million and Counting! *Physics Today*, *59*, 37.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force Concept Inventory. *The Physics Teacher*, *30*, 141-158.

Hume, G., Michael, J., Rovick, A., & Evens, M. (1996). Hinting as a tactic in one-on-one tutoring. *Journal of the Learning Sciences*, *5*(1), 23-49.

Hunt, E., & Minstrell, J. (1994). A Cognitive Approach to the Teaching of Physics. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice* (pp. 51-74). Cambridge, MA: MIT Press.

Jones, R. M., & Fleischman, E. S. (2001). Cascade explains and informs the utility of fading examples to problems. In *23rd Annual Conference of the Cognitive Science Society* (pp. 459-464). Lawrence Erlbaum Associates.

Jordan, P., Rose, C. P., & VanLehn, K. (2001). Tools for authoring tutorial dialogue knowledge. In J. D. Moore, C. Redfield, & W. L. Johnson (Eds.), *Artificial Intelligence in Education: AI-Ed in the Wired and Wireless future* (pp. 222-233). Washington, DC: IOS.

Jordan, P., Makatchev, M., Papuswamy, U., VanLehn, K., & Albacete, P. (2006). A Natural Language Tutorial Dialogue System for Physics. In G. Sutcliffe & R. Goebel (Eds.), *Proceedings of the 19th International FLAIRS conference* (p. 521). Menlo Park, CA: AAAI Press.

Katz, S. (2006, July). Post-practice dialogues in an intelligent tutoring system for college-level physics. . poster, Syracuse, NY.

Katz, S., Connely, J., & Wilson, C. (2007). Out of the lab and into the classroom: an evaluation of reflective dialogue in Andes. In R. Luckin, K. R. Koedinger, & J. Greer (Eds.), *Artificial Intelligence in Education* (Vol. 158, pp. 425-432). Amsterdam: IOS Press.

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*(1), 30-43.

Lakhani, K. R., & Hippel, E. V. (2003). How open source software works: "free" user-to-user assistance. *Research Policy*, *32*(6), 923-943. doi: 10.1016/S0048-7333(02)00095-1.

Larkin, J., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Models of competence in solving physics problems. *Cognitive Science*, *4*, 317-345.

Larkin, J., Reif, F., Carbonell, J., & Gugliotta, A. (1988). Fermi: A flexible expert reasoner with multi-domain inferencing. *Cognitive Science*, *12*(1), 101-138.

Liew, C. W., Shapiro, J. A., & Smith, D. E. (2004). Inferring the context for evaluating physics algebraic equations when scaffolding is removed. In *Proceedings of FLAIRS2004*.

Litman, D., Rose, C., Forbes-Riley, K., VanLehn, K., Bhembe, D., & Silliman, S. (2006). Spoken versus typed human and computer dialogue tutoring. *International Journal of Artificial Intelligence and Education*, *16*, 145-170.

Mayer, R. E. (1981). Frequency norms and structural analysis of algebra story problems into families, categories, and templates. *Instructional Science*, *10*(2), 135-175. doi: 10.1007/BF00132515.

Mazur, E. (1993). *Peer Instruction: A User's Manual*. Cambridge, MA: Harvard University Press.

Merrill, D. C., Reiser, B. J., Ranney, M., & Trafton, J. G. (1992). Effective tutoring techniques: A comparison of human tutors and intelligent tutoring systems. *The Journal of the Learning Sciences*, *2*(3), 277-306.

National Science Board. (2003). *The Science and Engineering Workforce: Realizing America's Potential*. Arlington, VA: National Science Foundation. Retrieved June 21, 2008, from http://www.nsf.gov/nsb/documents/2003/nsb0369/nsb0369.pdf.

National Science Board. (2007). *National Action Plan for Addressing the Critical Needs of the U.S. Science, Technology, Engineering, and Mathematics Education System*. Arlington, VA: National Science Foundation. Retrieved June 20, 2008, from http://www.nsf.gov/nsb/documents/2007/stem_action.pdf.

Nicaud, J., Bouhineau, D., Varlet, C., & Nguyen-Xuan, A. (1999). Towards a product for teaching formal algebra. In S. P. Lajoie & M. Vivet (Eds.), *Artificial Intelligence in Education* (pp. 207-214). Amsterdam: IOS Press.

Norman, D. A. (1981). Categorization of action slips. *Psychological Review*, *88*(1), 1-15.

Pascarella, A. M. (2002). *CAPA (Computer-Assisted Personalized Assignments) in a Large University Setting*. Doctoral Dissertation, University of Colorado.

Pascarella, A. M. (2004). The influence of web-based homework on quantitiatve problem-solving in university physics classes. In *National Association for Research in Science Teaching (NARST)*. Vancouver, BC, Canada.

Ploetzner, R., & VanLehn, K. (1997). The acquisition of informal physics knowledge during formal physics training. *Cognition and Instruction*, *15*(2), 169-206.

Priest, A. G., & Lindsay, R. O. (1992). New light on novice-expert differences in physics problem solving. *British Journal of Psychology*, *83*, 389-405.

Reif, F., & Scott, L. A. (1999). Teaching scientific thinking skills: Students and computers coaching each other. *American Journal of Physics*, *67*(9), 819-831.

Reimann, P., Wichmann, S., & Schult, T. (1993). A learning strategy model for worked-out examples. In P. Brna, S. Ohlsson, & H. Pain (Eds.), *Artificial Intelligence in*

Education: Proceedings of AI-ED93 (pp. 290-297). Charlottesville, VA: Association ofr Advancement of Computing in Education.

Renkl, A., & Atkinson, R. K. (2003). Structuring the transition from example study to problem solving in cognitive skill acquisition: A cognitive load perspective. *Educational Psychologist*, *38*(1), 15-22.

Rose, C. P., Di Eugenio, B., & Moore, J. D. (1999). A dialogue based tutoring system for basic electricity and electronics. In S. P. Lajoie & M. Vivet (Eds.), *Artificial Intelligence in Education: Open leraning Environments: New Computational Technologies to Support Learning, Exploration and Collaboration* (pp. 759-761). Amsterdam: IOS Press.

Rose, C. P., Jordan, P., Ringenberg, M., Siler, S., VanLehn, K., & Weinstein, A. (2001). Interactive conceptual tutoring in Atlas-Andes. In J. D. Moore, C. Redfield, & W. L. Johnson (Eds.), *Artificial Intelligence in Education: AI-Ed in the Wired and Wireless future* (pp. 256-266). Washington, DC: IOS.

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

Schwonke, R., Wittwer, J., Aleven, V., Salden, R., Kreig, C., & Renkl, A. (2007). Can tutored problem solving benefit from faded worked-out examples? In S. Vosniadou, D. Kayser, & A. Protopapas (Eds.), *Proceedings of the 2nd European Cognitive Science Conference* (pp. 59-64). Mahwah NJ: Erlbaum. Retrieved from http://www.cs.cmu.edu/~aleven/Papers/2007/Schwonke_ea_EuroCogSci2007.pdf .

Siler, S., Rose, C. P., Frost, T., VanLehn, K., & Koehler, P. (2002). Evaluating knowledge construction dialogues (KCDs) versus minilesson within Andes2 and alone. In C. P. Rose & V. Aleven (Eds.), *Workshop on dialogue-based tutoring at ITS 2002* (pp. 9-15). Biaritz, Spain.

Siler, S. A., & VanLehn, K. (2003). Accuracy of tutor's assessments of their students by tutoring context. In R. Alterman & D. Hirsch (Eds.), *Proceedings of the 25th Annual Meeting of the Cognitive Science Society* (pp. 1082-1087). Mahwah, NJ: Erlbaum.

Singley, M. K. (1990). The reification of goal structures in a calculus tutor: Effects on problem solving performance. *Interactive Learning Environments*, *1*, 102-123.

Van Heuvelen, A. (1991). Overview, Case Study Physics. *American Journal of Physics*, *59*(10), 898-907.

Van Heuvelen, A., & Zou, X. (2001). Multiple representations of work-energy processes. *American Journal of Physics*, *69*(2), 184-194.

VanLehn, K. (1999). Rule learning events in the acquisition of a complex skill: An evaluation of Cascade. *Journal of the Learning Sciences*, *8*(2), 179-221.

VanLehn, K., & Jones, R. M. (1993). Better learners use analogical problem solving sparingly. In P. E. Utgoff (Ed.), *Machine Learning: Proceedings of the Tenth Annual Conference* (pp. 338-345). San Mateo, CA: Morgan Kaufmann.

VanLehn, K., Siler, S., Murray, C., Yamauchi, T., & Baggett, W. B. (2003). Human tutoring: Why do only some events cause learning? *Cognition and Instruction*, *21*(3), 209-249.

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence and Education*, *16*.

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rose, C. P. (2005). When is reading just as effective as one-on-one interactive human tutoring? In *Proceedings of the 27th Annual Meeting of the Cognitive Science Society* (pp. 2259-2264). Mahwah, NJ: Erlbaum.

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rose, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science*, *31*(1), 3-62.

VanLehn, K., & Jones, R. M. (1993). Learning by explaining examples to oneself: A computational model. In S. Chipman & A. Meyrowitz (Eds.), *Cognitive Models of Complex Learning* (pp. 25-82). Boston, MA: Kluwer Academic Publishers.

VanLehn, K., Jones, R. M., & Chi, M. T. H. (1992). A model of the self-explanation effect. *The Journal of the Learning Sciences*, *2*(1), 1-59.

VanLehn, K., Jordan, P., & Litman, D. (2007). Developing pedagogically effective tutorial dialogue tactics: Experiments and a testbed. In . Farmington, PA.

VanLehn, K., Jordan, P., Rose, C. P., Bhembe, D., Bottner, M., Gaydos, A., et al. (2002). The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In S. A. Cerri, G. Gouarderes, & F. Paraguacu (Eds.), *Intelligent Tutoring Systems, 2002, 6th International Conference* (pp. 158-167). Berlin: Springer.

VanLehn, K., Koedinger, K. R., Skogsholm, A., Nwaigwe, A., Hausmann, R. G. M., Weinstein, A., et al. (2007). What's in a step? Toward general, abstract representations of tutoring system log data. In C. Conati & K. McCoy (Eds.), *Proceedings of User Modelling 2007* (pp. 465-469). Berlin: Springer-Verlag.

VanLehn, K., Lynch, C., Schultz, K., Shapiro, J. A., Shelby, R. H., Taylor, L., et al. (2005). The Andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence and Education*, *15*(3), 147-204.

VanLehn, K., & van de Sande, B. Expertise in elementary physics, and how to acquire it. In K. A. Ericsson (Ed.), *Development of professional expertise: Toward measurement of expert performance and design of optimal learning environments*.

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press.