The Design, Deployment and Evaluation of the AnimalWatch Intelligent Tutoring System

Paul Cohen*, Carole Beal*, Niall Adams**

* University of Southern California
**Imperial College London
Lessons About the Design, Deployment and Evaluation of the AnimalWatch ITS

• The state of mathematics education and the motivation for the AnimalWatch intelligent tutoring system (ITS)
• Design of the AnimalWatch ITS
• Issues in deploying ITSs
• Issues in evaluating ITSs
Poor Mathematics Education & Achievement

- In Los Angeles, only 36% of students are "proficient or better" on 2007 California State Test-Math
- Algebra 1 is required for graduation in many US schools
- High algebra failure rates in many urban schools; 40-80% in some CA districts; linked to high school dropout

- Similar challenges in many European countries
  Higher overall achievement (relative to US) but growing pool of students with low skills, and students not proficient in language of instruction
Contributing Factors

- Low proficiency with basic computation, fractions and pre-algebra (e.g., decimal-fraction equivalents, proportions)
- Lack of teacher preparation to teach mathematics
- High teacher turnover due to challenging classroom conditions, master teachers are retiring
- Large immigrant populations and difficulty of math instruction due to English comprehension
- High student-to-teacher ratios

Examples of problem solving errors in mathematics classes in Los Angeles.
Intelligent Tutoring Systems
(visit http://k12.usc.edu)

• AI-based Intelligent Tutoring Systems (ITS) such as AnimalWatch generally provide:
  – Individualized instruction for students
  – Skills practice
  – Real-time assessments for teachers

• Have been shown to be effective with adult learners in military, medical training apps, etc. and with K20 students teaching mathematics

• Recommended by National Mathematics Advisory Council report to U.S. Dept. of Education in May 2008
Design principles for ITSs for K12 mathematics

Connect Mathematics with Science

• The US National Council of Teachers of Mathematics recommends an integrated math-science curriculum - currently the subjects are taught independently.

• AnimalWatch organizes word problems around narratives dealing with endangered species. All the science content is authentic and checked with experts.
Design principles for ITSs for K12 mathematics

Problem-based Instruction

- Students are engaged by problems and by the interactive help they get solving them (a sequence of hints, worked examples, and for some topics, video clips of expert teachers)

1100 problems covering 30 math topics, indexed by difficulty.
Design principles for ITSs for K12 mathematics
One-on-one tutoring

- In AnimalWatch the next problem or hint *for a particular student* is selected by an expert system that takes into account
  - The student's level of mastery of each topic
  - The number of problems in each topic seen by the student
  - The difficulty of candidate problems relative to those the student has seen
  - How many hints the student has seen
  - Other versions of the problem selector additionally infer whether the student is guessing
Design principles for ITSs for K12 mathematics
Skills Practice with SkillBuilders

One reason for poor performance at algebra is poor preparation in basic arithmetic skills - fractions to percent, LCD, negatives, estimation, etc.

Fluent retrieval of math facts frees cognitive resources

Low difficulty & fast pace encourages repetition, builds proficiency, in an engaging alternative to word problems

Skill Builders

- **Fractions, Decimals and Percents**
  - Find the decimal and percent equivalents of fractions.
  - (E.g. \( \frac{1}{4} \) as a percentage is 25%)
  - start

- **Least Common Denominator**
  - Find the least common denominator (LCD) for two fractions.
  - (E.g. The LCD of \( \frac{1}{4} \) and \( \frac{1}{3} \) is 12)
  - start

- **Rounding**
  - Round very large & small numbers.
  - (E.g. Round 2,572,870 to the nearest thousands place: 2,573,000)
  - start

- **Estimate**
  - Estimate the approximate value of the sum of two numbers.
  - (E.g. 499 + 497 is approximately 1000)
  - start

- **Negative Integers**
  - Add & subtract negative integers.
  - (E.g. \(-3 + 5 = 2\))
  - start

- **Simple Algebra**
  - Practice one-step linear equations.
  - (E.g. \( 2x = 10, \ x = 5 \))
  - start
Principles for Deployment of ITSs

• Identify users: In-classroom at grade level, in-classroom review, after-school program, adult learners, community groups, etc.
  – AnimalWatch has been deployed with 6th graders, failing 9th graders, as review for USC "community of schools" pre-college program, with visually-impaired students, etc.

• Align ITS content with state (or national) standards
  – Teachers can't afford class time for subjects that won't be on annual tests. AnimalWatch is aligned with California State Curriculum.

• Assessment and learner tracking
  – One promise of ITSs is up-to-the-minute assessment at very fine granularity of each student's mastery of every topic. Teachers love this!
Principles for Deployment of ITSs
Assessment and Learner Tracking

AnimalWatch tracks

• Pre-post tests of target math skills
  • Computation
  • Fractions
  • Pre-algebra
• Math motivation
• Concrete-formal operational reasoning
• Spatial cognition
• Math fact proficiency
Principles for Deployment of ITSs
Be Prepared for Chaos

Student confrontations trigger Los Angeles Police Dept. arrival, school lockdown, missed AnimalWatch sessions
Principles for Deployment of ITs
Be Prepared for Chaos

Laptop carts were often unplugged overnight by custodial staff so computers were not ready for use when next school day started.
Principles for Deployment of ITSs
Be Prepared for Chaos

School networks were typically slow and became nearly unusable mid-day when students on lunch breaks began downloads of music & videos.
Principles for Deployment of ITSs
Be Prepared for Chaos

Access to school servers was not ideal! (No one at this school site had a key to the “cage”
Evaluating AnimalWatch
The Original Plan

- AnimalWatch was designed for 6th grade algebra
- We were approached by a charter school district to help their failing 9th graders
- Design: Pretest, then four sessions with AW, then Posttest
- Analysis: Test whether Posttest-Pretest difference is positive
- No control group because no teacher would participate in a control (no AnimalWatch) condition

- What actually happened was chaotic, from which we learned some principles for evaluating ITSs in difficult classrooms…
AWS Deployment: Student sample

• low-achieving student population: 40% failing math class; 80% not proficient on state exams

• high student attrition
  student drops out, leaves community, transfers to another school

• frequent new enrollments
  student appears at the school without notice or prior records, is assigned to a teacher on the spot

• many students not proficient in language of instruction (40-80%) yet state laws prohibit instruction in students’ primary language (Spanish)
Principles for Evaluating the Efficacy of ITGs

• Don't assume homogenous samples
  – The AnimalWatch study included 172 students of whom 88 spoke English as their native language (mostly African American) and 84 were English Language Learners (mostly Hispanic). Not equally distributed over the four classrooms in the study

• Design pretests and posttests to match the capabilities of the students (or risk ceiling or floor effects)
  – Over 80% of the students scored "below basic" or "far below basic" on the California Standards Test; 40% were failing their math class

• Assume there will be attrition
  – The sample was originally three times larger. One teacher opted out midway through, many students never took the posttest, many dropped out of school, etc.
Principles for Evaluating the Efficacy of ITs

• Don't assume a rational protocol
  – Because of scheduling constraints, availability of computers, lockdowns due to student brawls, etc., we couldn't guarantee that students got sessions with AnimalWatch at regular intervals

• Don't assume equal access. Consider dose-response design.
  – Because of chaotic conditions, some students got four sessions, some only one session. A standard pretest-posttest design wouldn't account for this. Our analysis became dose-response.
Results

Unique-AW-Problems: The number of AW problems a student solved (the "dose")
PreLO: Log(PretestCorrect / PretestIncorrect), i.e., initial math capability
PostLO: Log(PosttestCorrect / PosttestIncorrect) i.e., the "response"

PostLO = 0.66 PreLO + 0.21 Unique-AW-Problems

This regression model is highly significant (R^2 = .51, p<.0001). Coefficients are
standardized. Independent leverage plots shown above. Dose (work with AW)
clearly affects response (posttest performance) given pretest performance.
Conclusions

• Education is in crisis in the US and many other countries
• There are very few convincing evaluations of AI-based educational technology
• Some are nicely controlled lab studies. We've presented factors that affect the design, deployment and evaluation of ITSs for chaotic classrooms – the mode in large cities and poor areas
• We showed small, positive effects of AnimalWatch
• Further studies are needed
• Recruiting teachers and schools is hard...All four teachers in this study quit the profession before the end of their third year.