How to build tutoring systems that are almost as effective as human tutors?

Kurt VanLehn

School of Computing, Informatics and Decision Systems Engineering
Arizona State University
Outline

- Types of tutoring systems
- Step-based tutoring ≈ human tutoring
- How to build a step-based tutor
- Increasing their effectiveness
- Flame
Two major design dimensions

- **Personalization of assignments**
  - Non-adaptive
  - Competency gating
    » using sequestered assessments
    » one factor per module
  - Adaptive task selection
    » using embedded assessments
    » one factor per knowledge component

- **Granularity of feedback, hints & other interaction**
  - Assignment (e.g., conventional homework)
  - Answer (e.g., most regular tutoring systems)
  - Step (e.g., most Intelligent Tutoring Systems)
  - Sub-step (e.g., human tutors & some ITS)
Example: Pearson’s Mastering Physics

What tangential speed, \( v \), must the bob have so that it moves in a horizontal circle with the string always making an angle \( \theta \) from the vertical?

Express your answer in terms of some or all of the variables \( m, L, \) and \( \theta \), as well as the acceleration due to gravity \( g \).

\[ v = L \cdot g \cdot \sin(\theta) \cdot \tan(\theta) \]
Example: Andes Physics Tutor

**Personalization**
- Non-adaptive
  - Competency gating
  - Adaptive task selection

**Granularity**
- Answer
  - Step
  - Sub-step
Example: Cordillera Physics Tutor

- **Personalization**
  - Non-adaptive
  - Competency gating
  - Adaptive task selection

- **Granularity**
  - Answer
  - Step
  - Sub-step

---

A step

---

<table>
<thead>
<tr>
<th>Problem Statement</th>
<th>Dialog History</th>
<th>Variables</th>
<th>Equations</th>
</tr>
</thead>
<tbody>
<tr>
<td>You said: definition of kinetic energy</td>
<td>Tutor said: Okay. Please write the equation for how the definition of kinetic energy applies to this problem at T1. You said: value=(\text{ke1}=1/2m*v_1^2)</td>
<td>Tutor said: Now it is easy to calculate the magnitude of (v_1). The magnitude of (v_1) is 5.0 m/s. Tutor said: Before going on to the next step, let's think about the application of this equation. Tutor said: Can we infer the direction of the velocity of the rock at (T_1) from the rock's kinetic energy at (T_1)? You said: no, we cannot</td>
<td>Tutor said: Excellent! Please explain why. Only the magnitude of the velocity and not the direction of it is part of the definition of kinetic energy.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(m)</td>
<td>The mass of the rock is 0.60 kg</td>
</tr>
<tr>
<td>(v_0)</td>
<td>The velocity of the rock during (T_0) is 2.0 m/s at an unknown orientation.</td>
</tr>
<tr>
<td>(KE_0)</td>
<td>The kinetic energy of the rock at (T_0) is 1.20 J</td>
</tr>
<tr>
<td>(v_1)</td>
<td>The velocity of the rock during (T_1) is 5.0 m/s at an unknown orientation.</td>
</tr>
<tr>
<td>(KE_1)</td>
<td>The kinetic energy of the rock at (T_1) is 7.50 J</td>
</tr>
<tr>
<td>(W_{net01})</td>
<td>The work done on the rock</td>
</tr>
<tr>
<td>(TME_0)</td>
<td>The total mechanical energy of the system at (T_0)</td>
</tr>
<tr>
<td>(TME_1)</td>
<td>The total mechanical energy of the system at (T_1)</td>
</tr>
</tbody>
</table>
Example: Carnegie Learning’s Tutors

- **Personalization**
  - Non-adaptive
  - Competency gating
    - Adaptive task selection

- **Granularity**
  - Answer
    - Step
  - Sub-step
Carnegie Learning’s skillometer shows knowledge components & current competence

- Entering a given
- Identifying units
- Finding X, any form
- Writing expression
- Placing points
- Changing axis intervals
- Changing axis bounds
Example: Entity-relation Tutor

- **Personalization**
  - Non-adaptive
  - Competency gating
  - Adaptive task selection

- **Granularity**
  - Answer
  - Step
  - Sub-step
## Availability

<table>
<thead>
<tr>
<th></th>
<th>Non-adaptive</th>
<th>Competency gating</th>
<th>Adaptive task selection</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Answer-based feedback/hints</strong></td>
<td>Thousands</td>
<td>Hundreds</td>
<td>Few</td>
</tr>
<tr>
<td><strong>Step-based feedback/hints</strong></td>
<td>Hundreds (few on market)</td>
<td>Tens</td>
<td>Few</td>
</tr>
<tr>
<td><strong>Sub-step based feedback/hints</strong></td>
<td>Tens</td>
<td>None</td>
<td>None</td>
</tr>
</tbody>
</table>
Called CAI, CBT, CAL…

<table>
<thead>
<tr>
<th></th>
<th>Non-adaptive</th>
<th>Competency gating</th>
<th>Adaptive task selection</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Answer-based feedback/hints</strong></td>
<td>Thousands</td>
<td>Hundreds</td>
<td>Few</td>
</tr>
<tr>
<td><strong>Step-based feedback/hints</strong></td>
<td>Hundreds (few on market)</td>
<td>Tens</td>
<td>Few</td>
</tr>
<tr>
<td><strong>Sub-step based feedback/hints</strong></td>
<td>Tens</td>
<td>None</td>
<td>None</td>
</tr>
</tbody>
</table>
## Called Intelligent Tutoring Systems (ITS)

<table>
<thead>
<tr>
<th></th>
<th>Non-adaptive</th>
<th>Competency gating</th>
<th>Adaptive task selection</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Answer-based feedback/hints</strong></td>
<td>Thousands</td>
<td>Hundreds</td>
<td>Few</td>
</tr>
<tr>
<td></td>
<td>Hundreds (few on market)</td>
<td>Tens</td>
<td>Few</td>
</tr>
<tr>
<td><strong>Sub-step based feedback/hints</strong></td>
<td>Tens</td>
<td>None</td>
<td>None</td>
</tr>
</tbody>
</table>
Outline

- Types of tutoring systems
- Step-based tutoring ≈ human tutoring
- How to build a step-based tutor
- Increasing their effectiveness
- Flame
A widely held belief: Human tutors are much more effective than computer tutors.
A widely held belief: Human tutors are much more effective than computer tutors

<table>
<thead>
<tr>
<th>Effect Size</th>
<th>No tutoring</th>
<th>Computer Aided Instruction (CAI)</th>
<th>Intelligent tutoring systems (ITS)</th>
<th>ITS with natural language dialogue</th>
<th>Human tutors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bloom (1984)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anderson et al. (1995)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VanLehn et al. (2005)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Common belief: The finer the granularity, the more effective the tutoring

- Most ITS are step-based tutoring
- CAI is answer-based tutoring
Granularity of tutoring ≈ number of inferences (\(\rightarrow\)) between interactions

- **Answer-based tutoring (CAI)**

  ![Diagram showing the flow from problem to answer with intermediate steps]
Granularity of tutorial interaction ≈ number of inferences (→) between interactions

- **Answer-based tutoring (CAI)**
  - Problem
  - Student enters answer

- **Step-based tutoring (ITS with ordinary GUI)**
  - Problem
  - Student enters step
  - Student enters step
  - Last step
Granularity of tutorial interaction ≈ number of inferences (→) between interactions

◆ Answer-based tutoring (CAI)

<Diagram>

◆ Step-based tutoring (ITS with ordinary GUI)

<Diagram>

◆ Human tutoring

<Diagram>
Hypothesis: The smaller the grain size of interaction, the more effective the tutoring

- Because negative feedback is more effective
  - The shorter the chain of inferences, the easier to find the mistake in it

- Because hinting and prompting are more effective
  - The shorter the chain of inferences, the easier to infer them from a hint or prompt
Evidence for an interaction plateau

- 2 studies from my lab
- 3 studies from other labs
- A meta-analysis
Dialogue of Andes-Atlas

T: Here are a few things to keep in mind when computing the acceleration vector for a body at rest. Acceleration is change in what over time?

S: velocity

T: Right. If the velocity is not changing, what is the magnitude of the acceleration?

S: zero

T: Sounds good. ....

Text of Andes

Here are a few things to keep in mind when calculating acceleration for a body at rest. Acceleration is change in velocity over time. If velocity is not changing, then there is zero acceleration....
Results comparing Andes-Atlas to Andes

- Study 1: Andes-Atlas > Andes
  - but content not controlled properly
- Study 2 (N=26): Andes-Atlas ≈ Andes (p>.10)
- Study 3 (N=21): Andes-Atlas < Andes (p<.10, d=0.34)
- Study 4 (N=12): Andes-Atlas ≈ Andes (p>.10)

Conclusion: Substep tutoring is *not* more effective than step-based tutoring
The WHY2 studies
(VanLehn, Graesser et al., 2007, Cognitive Science)

5 conditions
- Human tutors
- Substep-based tutoring system
  » Why2-Atlas
  » Why2-AutoTutor (Graesser et al.)
- Step-based tutoring system
- Text

Procedure
- Pretraining
- Pre-test
- Training (~ 4 to 8 hours)
- Post-test
User interface for human tutoring and Why2-Atlas

Problem:
4. A rock is thrown straight upward with a velocity v. What is its acceleration at the top of its path? What is its velocity when it comes back to the starting point? Why?

Dialogue history

Student’s essay

Tutor: What is the velocity just before it hits the ground?
Student: That will depend on the time that the ball is in the air/distance ball traveled.
Tutor: How will it compare with the velocity with which it was thrown up?

Tell Tutor:
They will be equal. The height the ball reaches will depend on the initial velocity.

Enter your essay here:
At the very top of the path, where the velocity is equal to 0, the acceleration will be \(-9.8 \text{ m/(s}^2\text{)}\), the acceleration of gravity. It will no longer have any upward acceleration. When it comes back the its starting point (assuming that it is the ground) its velocity will be equal to 0 because its downward progress will be stopped by the ground.

Student’s turn in the dialogue
The sun exerts a gravitational force on the earth as the earth moves in its orbit around the sun. Does the earth pull equally on the sun? Explain why.

Tutor

Dialogue history

Task

Student types response
Only difference between tutoring conditions was contents of yellow box.
Human tutoring

- Tutor poses a WHY question
- Student response → analyzed as steps
- Tutor congratulates

Dialogue consisting of hints, analogies, reference to dialogue history...

Step is incorrect or missing
Why2-Atlas

Tutor poses a WHY question

Student response → analyzed as steps

Knowledge construction dialogue

Step is incorrect or missing

Tutor congratulates
Why2-AutoTutor

Tutor poses a WHY question

Student response → analyzed as steps

Step is incorrect or missing

Tutor congratulates

Hint, prompt, assert
A step-based tutor: A text explanation with same content

- Tutor poses a WHY question
- Student response → analyzed as steps
- Tutor congratulates

Text (the Why2-Atlas dialogue rewritten as a monologue)

Step is incorrect or missing
Experiments 1 & 2

No significant differences
Results from all 7 experiments

- Human tutoring
  = Substep-based tutoring systems
  = Step-based tutoring system
- Tutors > Textbook (no tutoring)
- Atlas (symbolic NLP) = AutoTutor (statistical NLP)

No significant differences

No tutoring
Reif & Scott (1999) also show human tutors = step-based tutoring

No significant differences

No significant differences
### Meta-analytic results for all possible pairwise comparisons (VanLehn, 2011)

<table>
<thead>
<tr>
<th>Tutoring type</th>
<th>vs. other tutoring type</th>
<th>Num. of effects</th>
<th>Mean effect</th>
<th>% reliable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Answer-based</td>
<td>no tutoring</td>
<td>165</td>
<td>0.31</td>
<td>40%</td>
</tr>
<tr>
<td>Step-based</td>
<td>no tutoring</td>
<td>28</td>
<td>0.76</td>
<td>68%</td>
</tr>
<tr>
<td>Substep-based</td>
<td>no tutoring</td>
<td>26</td>
<td>0.40</td>
<td>54%</td>
</tr>
<tr>
<td>Human</td>
<td>no tutoring</td>
<td>10</td>
<td>0.79</td>
<td>80%</td>
</tr>
<tr>
<td>Step-based</td>
<td>answer-based</td>
<td>2</td>
<td>0.40</td>
<td>50%</td>
</tr>
<tr>
<td>Substep-based</td>
<td>step-based</td>
<td>6</td>
<td>0.32</td>
<td>33%</td>
</tr>
<tr>
<td>Human</td>
<td>step-based</td>
<td>1</td>
<td>-0.04</td>
<td>0%</td>
</tr>
<tr>
<td>Substep-based</td>
<td>step-based</td>
<td>11</td>
<td>0.16</td>
<td>0%</td>
</tr>
<tr>
<td>Human</td>
<td>sub-step based</td>
<td>10</td>
<td>0.21</td>
<td>30%</td>
</tr>
<tr>
<td>Human</td>
<td>sub-step based</td>
<td>5</td>
<td>-0.12</td>
<td>0%</td>
</tr>
</tbody>
</table>
Graph of comparisons of 4 tutoring types vs. no tutoring

- No tutoring
- Answer-based (CAI)
- Step-based (ITS)
- Substep-based (ITS w/ NL)
- Human tutoring

Effect size
Graphing all 10 comparisons:
No tutor < CAI < ITS = ITS w/NL = human

Effect size

-0.5 0 0.5 1 1.5 2

No tutoring  Answer-based (CAI)  Step-based (ITS)  Substep-based (ITS w/ NL)  Human tutoring

vs. No tutoring
vs. Answer-based
vs. Step-based
vs. Substep-based
Graph of comparisons of 4 tutoring types vs. no tutoring

- **No tutoring**
- **Answer-based (CAI)**
- **Step-based (ITS)**
- **Substep-based (ITS w/ NL)**
- **Human tutoring**

Effect size

- Expected
- Observed
The interaction plateau hypothesis

- The smaller the grain size of interaction, the more effective the tutoring
  - Assignments < answers < steps
- But grain sizes less than steps are no more effective than steps
  - Steps = substeps = human
Limitations & caveats

◆ Task domain
  – Must allow computer tutoring
  – Only STEM; not language, music, sports...

◆ Normal learners
  – Not learning disabled
  – Prerequisite knowledge mastered

◆ Human tutors must teach same content as computer tutors
  – Only the type of tutoring (human, ITS, CAI) varies

◆ One-on-one tutoring
Outline

- Types of tutoring systems
- Step-based tutoring ≈ human tutoring
- How to build a step-based tutor
- Increasing their effectiveness
- Flame
Main modules of a non-adaptive step-based tutoring system
Main modules of an adaptive step-based tutoring system

- Student interface
- Step analyzer
- Feedback & hint generator
- Task selector
- Assessor (contains learner model)

Task loop

Step loop
Main types of step analyzers

Three main methods for generating ideal steps

- *Model tracing*: One expert system that can solve all problems in all ways
- *Example tracing*: For each problem, all acceptable solutions
- *Constraint-based*: Example + recognizers of bad steps + recognizers of steps equivalent to example’s steps

Comparing student and ideal steps

- Trivial if steps are menu choices, numbers, short texts
- Harder if steps are math, logic, chemistry, programming
- Use statistical NLP for essays, long explanations
- Use probabilistic everything for gestures
Outline

- Types of tutoring systems
- Step-based tutoring ≈ human tutoring
- How to build a step-based tutor
- Increasing their effectiveness
- Flame
The details can make a huge difference. How can we get them right?

- Called A/B testing in the game industry
- During example-based tutoring, when should the tutor *tell* the student an inference vs. *elicit* it from the student?
- Can machine-learned policies improve the *tell vs. elicit* decision?
- Min Chi’s Ph. D. thesis
**T:** Next we will calculate the rock's instantaneous velocity at T1.

**Elicit**

**T:** What principle should we apply?

**S:** Definition of Kinetic Energy

**Tell**

**T:** To calculate the rock's instantaneous velocity at T1, we will apply the definition of kinetic energy again.

**Tell**

**T:** Okay, let me just write the equation: \( ke1 = \frac{1}{2} m v1^2 \)

**Elicit**

**T:** Please write the equation for the application of the definition of kinetic energy at time T1.

**S:** \( ke1 = \frac{1}{2} m v1^2 \).
## 5-Stage Procedure

<table>
<thead>
<tr>
<th>Stage</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage 1</td>
<td>Study: 64 students using random policy.</td>
</tr>
<tr>
<td>Stage 2</td>
<td>Calculate Sub-optimal policy.</td>
</tr>
<tr>
<td>Stage 3</td>
<td>Study: 37 students using Sub-optimal policy</td>
</tr>
<tr>
<td>Stage 4</td>
<td>Calculate Enhancing &amp; Diminishing policies.</td>
</tr>
<tr>
<td>Stage 5</td>
<td>Study: 29 students using Enhancing policy vs. 28 students using Diminishing policy</td>
</tr>
</tbody>
</table>

Diminishing policy is calculated to decrease learning. Other policies are calculated to increase learning.
Calculated policies are composed of many rules, such as:

\[
\text{If problem: } \text{difficult} \\
\text{And last tutor action: } \text{tell} \\
\text{And student performance: } \text{high} \\
\text{And duration since last mention of the current principle } \geq 50 \text{ sec}
\]

Machine learner selected features in left side of rule from 50 possible features defined by humans
Results
(NLG = normalized learning gain)

Enhancing > everything else, which were about the same

$p = 0.77.$

$p = 0.02$

$p < 0.001$
Conclusions’ from Min Chi’s thesis

- Details do matter e.g., the Tell vs. Elicit decision
- Improved policies for Tell vs. Elicit can be induced from *modest amounts of data*
  - 103 students
- Induced policies can have a large effect on learning gains (d=0.8).
- Developers should do many such A/B studies
Overall conclusion: We need to use more step-based tutors

<table>
<thead>
<tr>
<th></th>
<th>Non-adaptive</th>
<th>Competency gating</th>
<th>Adaptive task selection</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Answer-based</strong></td>
<td>Thousands</td>
<td>Hundreds</td>
<td>Few</td>
</tr>
<tr>
<td>feedback/hints</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Step-based</strong></td>
<td>Hundreds (few on market)</td>
<td>Tens</td>
<td>Few</td>
</tr>
<tr>
<td>feedback/hints</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sub-step</strong></td>
<td>Tens</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>based feedback/hints</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Outline

- Types of tutoring systems
- Step-based tutoring ≈ human tutoring
- How to build a step-based tutor
- Increasing their effectiveness
- Flame
Why are there so few step-based tutoring systems?

- K-12 curriculum and standardized tests have evolved to favor answer-based tasks
- K-12 instructors do not view homework as the problem area; it’s classroom time that concerns them.
- Instructors need to share knowledge, policies and authority with a tutoring system
Why are competency-gated tutoring systems so rare?

- Schools are time-gated, not competency-gated
- Difficulty enforcing deadlines
- Grading based on time-to-mastery may be pointless and harmful.
Recommendation for instructors

- Use competency-gated tutoring system
  - Flip: Videos/reading at home. Exercises in class.
  - Half group work (paper?) and half individual work (tutor)
  - Noisy study halls instead of lecture halls
  - Deadlines & exams for core. Badges for enrichment.

- Use a step-based tutoring system
  - Buy one if you can
  - If you build one, use example-tracing first
  - If you will use it repeatedly, plan on A/B testing
Recommendations for parents

- Human tutors ≈ step-based tutoring systems
- If you can do the task, then you can tutor the task
  - Do not lecture/demo!
  - Be step-based.
Thank you!
The meta-analysis

Why2 experiments

Andes, the physics tutor

Andes-Cordillera study
- in prep
Andes-Atlas studies

Machine learning of Cordillera policies

Teaching a meta-cognitive strategy (MEA)