What does the Interaction Plateau imply for Lifelong Learning Companions?

Kurt VanLehn
School of Computing and Informatics
Arizona State University
Outline

◆ Terminology
  – Answer-based tutoring
  – Step-based tutoring, with remediation via
    » Hint sequences
    » Dialogues

◆ Hypotheses

◆ Evidence

◆ Implications for Lifelong Learning Companions
Answer-based tutoring systems (= CAI, LMS) have a task loop

1. Someone selects a task
2. Student (re-)enters answer
3. Tutor gives hints
4. Tutor congratulates
5. If incorrect, repeat from step 1
Step-based tutoring systems (= ITS) also have a step loop

- Someone selects a task
- Student (re-)enters step
- Tutor gives hints
- Tutor congratulates
- Incorrect
Andes user interface

Read a physics problem

Draw vectors

Type in equations

Type in answer

A 2000 kg car in neutral at the top of a 20.0 deg inclined driveway 20.0 m long slips its parking brake and rolls down. If we ignore friction and drag, what would the magnitude of the velocity of the car be when it hits the garage door?

Answer:

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

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

1. \( mc = 2000 \text{ kg} \)
2. \( d = 20.0 \text{ m} \)
3. \( Fw_y = mc \times g \)
Andes feedback and hints

A 2000.0 kg car in neutral at the top of a 20.0 deg inclined driveway 20.0 m long slips its parking brake and rolls down.

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

Answer:

- **Green means correct**
- **Red means incorrect**

**Hints**

1. Now that you have stated all of the given information, you should start on the major principles. What quantity is the problem seeking?
2. The magnitude of the instantaneous velocity of the car at time T1
3. Yep. What is the first principle application that you would like to work on? Hint: this principle application will usually be one that mentions the sought quantity explicitly. Therefore its equation may contain the sought quantity that the problem seeks.
Andes remedies incorrect steps with hint sequences

Someone selects a task

Student (re-)enters step

Tutor congratulates

The goal hint is sometimes a menu-based dialogue

<step turns red>
<goal hint>
<teaching hint>
<bottom out hint>

Incorrect
Problem: A rock climber is currently on the side of a cliff 67 feet off the ground. She can climb on average about two and one-half feet per minute.

1. When will she be 92 feet off the ground?
2. In twenty minutes, how many feet above the ground will she be?
3. In 75 seconds, how far above the ground will she be?
4. Ten minutes ago, how far above the ground would she have been?

Step: Enter an equation

Step: Fill in a cell

Step: Define an axis

Step: Divide both sides

Step: Plot a point
The database that the problem refers to:

<table>
<thead>
<tr>
<th>Table Name</th>
<th>Attribute List</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIRECTOR</td>
<td>number, lname, fname, born, died</td>
</tr>
<tr>
<td>MOVIE</td>
<td>number, title, type, aamazon, aawon, year, critics, director</td>
</tr>
<tr>
<td>STAR</td>
<td>lname, fname, number, born, died, city</td>
</tr>
<tr>
<td>CUSTOMER</td>
<td>lname, fname, number, address, rentals, bonus, jdate</td>
</tr>
<tr>
<td>TAPE</td>
<td>code, movie, pdate, times, customer, hiredate</td>
</tr>
<tr>
<td>STARS_IN</td>
<td>movie, star, role</td>
</tr>
</tbody>
</table>
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.
Steps are unordered and in natural language

Why do most computers have a disk drive? Why can’t they have only RAM?

1. RAM’s content disappears when power quits, but disk content persist.
2. RAM is usually holds less information than disk
3. RAM takes battery power, so larger RAM takes more power
4. Certain information, e.g., operating system and user files, must be stored permanently.
Why do most computers have a disk drive? Why can’t they have only RAM?

S: RAM is too small. Only the disk is big enough.

T: True. But if it had a lot of RAM, would that suffice?

S: The battery would run out too fast.

T: Excellent. What else?

S: That’s it.

T: What if the battery dies?

S: Oh. The RAM dies.

T: Anything wrong with that?

S: You lose your files.

T: Besides the user’s files, what else would be lost?

S: Beats me.

T: The operating system.
AutoTutor

Someone selects a task

Student (re-)enters step

Tutor congratulates

Incorrect or missing

Hint 1: <hint>
Hint 2: <prompt>
Hint 3: <assert>
Student’s screen for human tutoring and Why2/Atlas

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?

Student’s turn in the dialogue:

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?

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

Problem:

Dialogue history

Student’s essay

Send Essay or Message
Someone selects a task

Student (re-)enters step

Tutor congratulates

A Knowledge Construction Dialogue (KCD)

T: What’s its velocity when it hits the ground?
S: zero?
T: I meant just before it hits the ground.
S: Oh. 9.8 m/s?
T: No. Let’s figure this out. What are…
Outline

- Terminology
  - Answer-based tutoring
  - Step-based tutoring, with remediation via
    » Hint sequences
    » Dialogues

- Hypotheses
  Next

- Evidence

- Implications for lifelong learning companions
Hypothesized ranking of tutoring, most effective first

A. Expert human tutors
B. Ordinary human tutors
C. Natural language tutoring systems (i.e., step-based tutoring systems with dialogue as remediation)
D. Step-based tutoring systems with hint sequences as remediation
E. Answer-based tutoring systems
F. No tutoring
Hypothesized effect sizes

Learning gains

0
0.5
1
1.5
2
2.5

Expert human tutors
Ordinary human tutors
Step-based tutoring
Answer-based tutoring
No tutoring
Hypothesized effect sizes

Bloom’s (1984) 2-sigma: 4 weeks of human tutoring vs. classroom

Learning gains

Expert human tutors
Ordinary human tutors
Step-based tutoring
Answer-based tutoring
No tutoring

Classroom
Hypothesized effect sizes

Kulik (1984) meta-analysis of CAI vs. classroom → 0.4 sigma

Learning gains

Expert human tutors
Ordinary human tutors
Step-based tutoring
Answer-based tutoring
No tutoring

Classroom
Many intelligent tutoring systems: e.g., Andes (VanLehn et al, 2005), Carnegie Learning’s tutors...
My main claim:  
There is an interaction plateau

<table>
<thead>
<tr>
<th>Learning gains</th>
<th>Expected</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Expert human tutoring
- Ordinary human tutoring
- Answer-based tutoring
- Step-based tutoring
- No tutoring

Expected gains: 0, 0.5, 1, 1.5, 2, 2.5
Observed gains: 0, 0.5, 1, 1.5, 2, 2.5
Outline

- Terminology
- Hypothesis
- Evidence
- Implications for Lifelong Learning Companions
Experiments had 6 conditions
(VanLehn, Graesser et al., 2007)

- **Expert Human tutors**
  - Typed
  - Spoken

- **Natural language tutoring systems**
  - Why2-AutoTutor (Graesser et al.)
  - Why2-Atlas (Jordan, Rosé, VanLehn et al.)

- **Step-based tutoring system**
  - Canned text remediation

- **No tutoring**
  - Textbook
Only difference between conditions was contents of yellow box

1. Someone selects a task
2. Student (re-)enters step
3. Tutor congratulates

Incorrect or missing
Human tutors

Someone selects a task

Student (re-)enters step

Tutor congratulates

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

Incorrect or missing
Why2-Atlas

Someone selects a task

Student (re-)enters step

Tutor congratulates

Incorrect or missing

Knowledge construction dialogue
Why2-AutoTutor

1. Someone selects a task
2. Student (re-)enters step
3. Tutor congratulates
4. Hint, prompt, assert

Flow Diagram:
- Someone selects a task
- Student (re-)enters step
- Tutor congratulates
- Incorrect or missing
- Hint, prompt, assert
Canned text

Someone selects a task

Student (re-)enters step

Tutor congratulates

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

Incorrect or missing
Experiments 1 & 2
(VanLehn, Graesser et al., 2007)

No significant differences

Adjusted post-test scores

- Read-only textbook studying
- Canned-text
- Why2-AutoTutor
- Why2-Atlas
- Human tutoring
Results from all 7 experiments (VanLehn, Graesser et al., 2007)

◆ Why2: Atlas = AutoTutor
◆ Why2 > Textbook
  – No essays
  – Content differences
◆ Human tutoring = Why2 = Canned text remediation
  – Exception: When pre-physics students worked with instruction authored for post-physics students, then Human tutoring > Canned text remediation
Why2 results support interaction plateau hypothesis

Expected vs. Observed Learning gains

- No tutoring
- Step-based tutoring
- Expert human tutoring
Other evidence for the interaction plateau (Evens & Michael, 2006)

A step-based tutoring system with text remediation

No significant differences

No tutoring
Circsim results support interaction plateau hypothesis

Expected

Observed

Learning gains

No tutoring
Step-based tutoring
Nat. lang. tutoring
Expert human tu...

34
Other evidence for the interaction plateau (Reif & Scott, 1999)

No significant differences
Reif & Scott results support interaction plateau hypothesis

- **Expected**
- **Observed**

<table>
<thead>
<tr>
<th>Step-based tutoring</th>
<th>Expert human tutoring</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning gains</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1.5</td>
<td>1.5</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2.5</td>
<td>2.5</td>
</tr>
</tbody>
</table>

No tutoring

No significant differences
Katz et al. results support the interaction plateau hypothesis

![Graph showing learning gains comparison]

- **Expected**
- **Observed**

<table>
<thead>
<tr>
<th>Tutor Type</th>
<th>Learning Gains</th>
</tr>
</thead>
<tbody>
<tr>
<td>No tutoring</td>
<td>0.5</td>
</tr>
<tr>
<td>Step-based tutoring</td>
<td>1.5</td>
</tr>
<tr>
<td>Ordinary human</td>
<td>2.0</td>
</tr>
</tbody>
</table>
### Evidence: 43 comparisons

*(VanLehn, submitted)*

<table>
<thead>
<tr>
<th>Studies</th>
<th>No tutoring</th>
<th>Answer-based tutoring</th>
<th>Step-based tutoring</th>
<th>Human tutoring</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Solve w/o feedback</td>
<td>Reading / studying</td>
<td>Copy example</td>
<td>Text remediation</td>
</tr>
<tr>
<td>1 Evens &amp; Michaels, 2006</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Evens &amp; Michaels, 2006</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Evens &amp; Michaels, 2006</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 VanLehn, Graesser et al., 2007</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 VanLehn, Graesser et al., 2007</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 VanLehn, Graesser et al., 2007</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 Reif &amp; Scott 1999</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 Reif &amp; Scott 1999</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9 Rose, Alevin et al., 2005</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 Fossati et al., 2008</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11 Katz, Allbritton &amp; Connelly, 2003</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12 Chae, Kim &amp; Glass, 2005</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13 Johnson &amp; Johnson, 1992</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14 Rose et al., 2001</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15 Utman et al., 2006</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16 Bloom, 1984</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17 Merrill et al., 1995</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18 Azvedo et al., 2006</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19 Witherspoon et al., 2007</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21 Swanson, 1992</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22 Chi, Roy &amp; Hausmann, 2008</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23 Chi, Roy &amp; Hausmann, 2008</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24 Weerasinghe &amp; Mitrovic, 2006</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25 Sier, Rose et al., 2002</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26 Person et al.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27 Graesser et al, 2003</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28 Arnott, Hastings &amp; Allbritton, in press</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29 Craig et al., 2004; 2006</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30 Craig et al., 2006 with questions</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31 Lane &amp; VanLehn, 2005</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32 Hefferman et al., 2008</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33 Mendicino et al., in press</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34 Razzaq et al., 2008</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35 Moreno, Mayer et al., 2001</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>36 VanLehn et al., 2005</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>37 Corbett &amp; Anderson, 2001</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38 Anderson et al, 1995</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>39 Koedinger et al., 1997</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40 Roberts, Pioch &amp; Ferguson, 2000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>41 Mitrovic &amp; Ohlsson, 1999</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>42 Suraweera &amp; Mitrovic, 2004</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>43 Charney, Reder &amp; Kusbit, 1990; 1986</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
My main claim (again):
There is an interaction plateau

![Graph showing learning gains with different tutoring methods]

- Expected
- Observed

Learning gains

- Expert human tutoring
- Ordinary human tutoring
- Answer-based tutoring
- Step-based tutoring
- No tutoring
Conclusion: How tutors present the remediation has no effect on learning gains

Someone selects a task

Student (re-)enters step

Tutor congratulates

Incorrect or missing

**Content** here matters. **Form** (e.g., dialogue, hint sequence, text) doesn’t matter
Why is there an interaction plateau?

- Instructors design steps so that an ideal student can just bridge each one. Thus, simple remediation suffices if the student can’t enact a step.
- When a student has failed to do a step, the student is motivated to learn from the remedial text, hint sequence or dialogue. Thus, more interactivity yields no added value.
Step-based ITS should be more effective than expert human tutors

◆ Inner step loop
  – Although ITS’s remediations are simpler (e.g., mere hint sequences), they are just as effective
  – ITS makes fewer mistakes interpreting steps

◆ Outer task loop
  – ITS can do more accurate, deeper assessments
  – ITS can accurately index a larger library of tasks
Outline

- Terminology
- Hypothesis
- Evidence
- Implications for Lifelong Learning Companions
If we duck the NL problem, then…

- E.g., avoid tasks that require NL
A lifelong learning companion (LLC) could be just a fancy LMS!

Someone selects a task

Student, LLC, peers work on the task

LLC rewards student
LLC supports many task types (feasible, sort of)

- Someone selects a task
- Student, LLC, peers work on the task
- LLC rewards student

- Step-based tutoring
- Text, videos, etc.
- CSCL
- Projects
- Hands on
- Etc
Task selection gives learner more control (feasible)

- Someone selects a task
- Student, LLC, peers work on the task
- LLC rewards student

- Pre-requisites
- Mastery learning, practice scheduling
- Possible futures
- Interests
- Learning community’s (or the Army’s) needs
LLC gives larger rewards (feasible?)

Someone selects a task

Student, LLC, peers work on the task

LLC rewards student

- Leveling up
- Access to interesting people, topics, …
- Real certificates
- Real money
BUT: LLC must understand tasks well enough to select, support, reward

◆ LMS solution
  – Meta data for each task

◆ ITS solution
  – Solution graph (like Collagen’s HTN) for each task
  – Perhaps generated by a problem solver/planner

◆ How to understand other types of tasks?
  – CSCL tasks
  – Projects
  – Explorations…
Biggest problem is knowledge "engineering" bottleneck

- Lifelong = 80 years = 30K days @ 4 hrs/day = 120K hours = 200K tasks
- Andes has ~500 tasks, covers 2 semesters
- Open ended & growing library
- Must automate the task analysis
  - LSA assigns metadata?
  - Social computing?
Summary

- 200K analyzed tasks (hard)
- Novel task selection, reward (feasible)
- NL for tasks that need them (optional)
- A lifelong learning companion
Back up slides are next
When should natural language be used in step-based tutoring systems?

◆ Only when there is no alternative
  – Spoken commands of COVE (Roberts, 2001)
  – Spoken commands of SCoT (Pon-Barry et al., 2006)
  – Tactical Iraqi (Johnson et al.)
  – Predator & Aegis team training
  – Essays of Why2?
  – Many others…
Why did Bloom (1984) observe 2.0 effect size?

◆ Three conditions
  – Human tutors with mastery learning with a 90% threshold → 2.0 sigma
  – Classroom with mastery learning with an 80% threshold → 1.0 sigma
  – Classroom

◆ My interpretation:
  – It’s the mastery learning thresholds, not the human tutoring
Why does the public think expert human tutors are gods?

- Human tutors > step-based tutoring system when the material is so far over the student’s head that they rarely understand the text/hints used for remediation (VanLehn et al., 2007)

- Step-based tutoring systems are not yet common