An Introduction to Intelligent Tutoring Systems (ITS)
(a.k.a. cyber-tutoring, digital tutors, ICAI)

Kurt VanLehn
Arizona State University
Tempe, AZ
Soller et al. (2005) claim edutech innovation is either:

- **Structural**: Changes the lesson plan, content and activities
- **Regulative**: Adds a regulative (i.e., cybernetic; feedback) loop:
  - **Sense** the students’ performance
  - **Compare** the students’ performance to **Expectations**
  - **Act** to decrease $\Delta$ between the students’ actual and expected performance

---

Main components of an ITS
Viewed as a regulative loop

Actual student performance

Compare

Expected student performance

Tutor’s actions to decrease difference
To design an ITS, choose at least one from each column

<table>
<thead>
<tr>
<th>Actual student performance</th>
<th>Expected student performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sources:</td>
<td>Sources:</td>
</tr>
<tr>
<td>- Answers to questions</td>
<td>- Expert authors</td>
</tr>
<tr>
<td>- Essays</td>
<td>- Algorithms</td>
</tr>
<tr>
<td>- Actions in a game</td>
<td>- Etc.</td>
</tr>
<tr>
<td>- Etc.</td>
<td>- Etc.</td>
</tr>
</tbody>
</table>

Compare

Tutor’s actions to decrease difference

Action types:
- Give feedback
- Choose next task
- Etc.
I will make 4 main points

1. Many sources are feasible
2. More frequent data are better, up to a point.
3. Human expert authors are (still) the main source.
4. Only three tutor action types have strong evidence of effectiveness.

Compare

Actual student performance

Expected student performance

Tutor’s actions to decrease difference
Main sources of student performance data

- Answer-based
  - Tutor assigns task, then student (eventually) enters a short answer e.g., multiple choice, number, drag & drop...

- Step-based
  - Tutor assigns task, then student makes many actions observed by the tutor (steps).

- Spoken student discussions
  - Tutor assigns task, then a small group of students discuss orally.
Next 7 slides are examples of step-based tutors’ user interfaces

- **Step-based**
  - Tutor assigns task, then student makes many actions observed by the tutor (steps).

Now feasible
An editor for solving physics problems

An editor for constructing concept maps

An editor for complex math problem solving
Tutor-student dialogue

Exellent! Please explain why.

Only the magnitude of the velocity and not the direction of it is part of the definition of kinetic energy.

An editor for drawing explanations

A multiplayer game

Diagnosing an epidemic

Step

How close are tutors to understanding unconstrained speech?

• **Answer-based**
  - Tutor assigns task, then student (eventually) enters a short answer e.g., multiple choice, number, drag & drop...

• **Step-based**
  - Tutor assigns task, then student makes many actions observed by the tutor (steps).

• **Spoken student discussions**
  - Tutor assigns task, then a small group of students discuss orally.
Tutor’s **can** understand constrained speech

Radio operator (RTO) practices calling for artillery fire

1. RTO: steel one niner this is gator niner one adjust fire polar over
2. FSO: gator nine one this is steel one nine adjust fire polar out
3. RTO: direction five niner four zero distance four eight zero over
4. FSO: direction five nine four zero distance four eight zero over

http://www.netc.navy.mil/centers/swos/Simulators.htm
Tutors can understand short answers to their questions

Tutors can understand affect & collaboration in spoken conversations in lab settings

Tutors cannot yet understand the content of unconstrained conversation between students, even in lab settings.
Main points: Transition slide

1. Many sources are feasible.
2. More frequent data are better, up to a point.
3. Human expert authors are (still) the main source.
4. Only three tutor action types have strong evidence of effectiveness.

Actual student performance

Expected student performance

Compare

Tutor’s actions to decrease difference

Done

Next
More frequent tutor-student interactions foster more learning, up to a point

<table>
<thead>
<tr>
<th>Tutoring type</th>
<th>vs. other tutoring type</th>
<th>Num. of effects</th>
<th>Mean effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Answer-based</td>
<td>no tutoring</td>
<td>165</td>
<td>0.31</td>
</tr>
<tr>
<td>Step-based</td>
<td></td>
<td>28</td>
<td>0.76</td>
</tr>
<tr>
<td>Human</td>
<td></td>
<td>10</td>
<td>0.79</td>
</tr>
</tbody>
</table>

• Answer-based > no tutoring by 0.30
• Step-based tutoring > answer-based by 0.45
• Human tutoring = step-based tutoring

More frequent tutor-student interactions foster more learning, up to a point

<table>
<thead>
<tr>
<th>Tutoring type</th>
<th>vs. other tutoring type</th>
<th>Num. of effects</th>
<th>Mean effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Answer-based</td>
<td>no tutoring</td>
<td>165</td>
<td>0.31</td>
</tr>
<tr>
<td>Step-based</td>
<td></td>
<td>28</td>
<td>0.76</td>
</tr>
<tr>
<td>Human</td>
<td></td>
<td>10</td>
<td>0.79</td>
</tr>
<tr>
<td>Step-based</td>
<td>answer-based</td>
<td>2</td>
<td>0.40</td>
</tr>
<tr>
<td>Human</td>
<td>answer-based</td>
<td>1</td>
<td>-0.04</td>
</tr>
<tr>
<td>Human</td>
<td>step-based</td>
<td>10</td>
<td>0.21</td>
</tr>
</tbody>
</table>

- Answer-based > no tutoring by 0.30
- Step-based tutoring > answer-based by 0.45
- Human tutoring = step-based tutoring

Main points: Transition slide

1. Many sources are feasible
2. More frequent data are better, up to a point.
3. Human expert authors are (still) the main source.
4. Only three tutor action types have strong evidence of effectiveness.

Actual student performance

Expected student performance

Tutor’s actions to decrease difference

Done

Next
Authoring:

• A human author invents the task
• Expected student performance on it = set of steps
  • Each step is marked as correct vs. incorrect
  • May also be marked for concepts & misconceptions
• Sources of expected performances (steps)
  • Human author performs the task in all ways
  • Students mark each other’s performances
  • Algorithm performs the task in all ways
  • Human authors one performance; algorithm generates all equivalents
  • Algorithm clusters student performances; human marks the prototype of each cluster

Well-defined task domains only
Main points: Transition slide

1. Many sources are feasible
2. More frequent data are better, up to a point.
3. Human expert authors are (still) the main source.
4. Only three tutor action types have strong evidence of effectiveness.
Common activities in classes.

- Reading & watching videos
- Whole class lectures & discussions
- Assessments (i.e., tests)
- Individual practice
- Small group work
- Projects
- Field trips

ITS are feasible

Next
Strong evidence that **adaptive assessment** is more effective

- After the student enters the answer to a task
  - System updates its estimate of the student’s mastery
  - System choose task that will maximize information gain
  - System present the task to the student

- Effectiveness
  - Validity – same as convention assessment
  - Reliability – same or better
  - Efficiency – better or same

- Widely used

Likely that embedded assessment is more effective

- The ITS updates estimates of student’s competence as the student gets feedback, hints, etc.
  - Assessing a moving target
- Practical advantages
  - No time wasted on testing
  - No test anxiety
  - No make-up tests
  - No test security issues
- Effectiveness
  - Reliability – excellent, but not clear how to compare
  - Validity – few studies

Strong evidence that mastery learning increases learning

- Mastery learning (also called Gating)
  - *Conventional* assessment: If you fail the test at the end of the module, you must study the module again and try the test again.
  - *Embedded* assessment: Keeping doing tasks until the ITS says you can go to the next module.

- Many studies, with & without ITS
  - Across 108 studies, effect size = 0.52

Strong evidence that feedback & hints increase learning

• As mentioned earlier
  • Answer-based vs. no-tutoring: 0.31 effect size
  • Step-based vs. no-tutoring: 0.76
  • Human tutors vs. no-tutoring: 0.79

• Most recent meta-evaluations
  • Answer based (CAI) vs. baseline: 0.38
  • ITS vs. baseline: 0.66
  • Human tutors vs. baseline: 0.40

Adaptive **task selection:**
Weak evidence or effect

- Choose tasks to match the student’s learning style
  - No evidence (yet) of effectiveness

Adaptive task selection: Weak evidence or effect

• Choose tasks to match the student’s learning style
  • No evidence (yet) of effectiveness

• Let the learner, not the system, choose tasks
  • Very small effect size

Adaptive task selection: Weak evidence or effect

• Choose tasks to match the student’s learning style
  • No evidence (yet) of effectiveness

• Let the learner, not the system, choose tasks
  • Very small effect size

• Space repeated tasks far apart
  • Strong evidence but only for memorization

Adaptive task selection: Weak evidence or effect

• Choose tasks to match the student’s learning style
  • No evidence (yet) of effectiveness

• Let the learner, not the system, choose tasks
  • Very small effect size

• Space repeated tasks far apart
  • Strong evidence but only for memorization

• Task difficulty matches student’s competence
  • No studies apart from mastery learning?
Adaptive task selection: Weak evidence or effect

• Choose tasks to match the student’s learning style
  • No evidence (yet) of effectiveness
• Let the learner, not the system, choose tasks
  • Very small effect size
• Space repeated tasks far apart
  • Strong evidence but only for memorization
• Task difficulty matches student’s competence
  • No studies apart from mastery learning?
• Choose tasks with a few unmasted topics
  • Just one study?

ITS impact on **small group work**: Weak evidence

- Feedback and hints
  - Most studies focus on increasing collaboration
  - Few studies measure learning
- Selecting group members
  - Few studies measure learning
ITS impact on teachers: Weak evidence

- Freeing teachers to help neediest students
- Teachers can focus on reviewing problematic tasks
- Use of dashboards during class

Progress bars help teachers decide when activity is done

Most common issue. Consider pausing class to discuss it

FACT recommends visiting group 2
1. Many sources are feasible.

2. More frequent data are better, up to a point.

3. Human expert authors are (still) the main source.

Tutor’s actions to decrease difference:
- Strong evidence
  - Feedback & hints
  - Mastery learning
  - Adaptive assessment
- Likely
  - Embedded assessment
- Weak evidence or effect
  - Adaptive task selection
  - Impact on small groups
  - Impact on teachers
A tutor-student dialogue that starts with an essay question:

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.

No, the sun is much more massive than the earth, so it pulls harder. That is why the earth orbits the sun and not vice versa.

Student’s initial answer is short essay, analyzed into propositions (steps).

Tutor’s initial question

Subsequent dialogue