An Introduction to Intelligent Tutoring Systems (ITS)

(a.k.a. cyber-tutoring, digital tutors, ICAI)

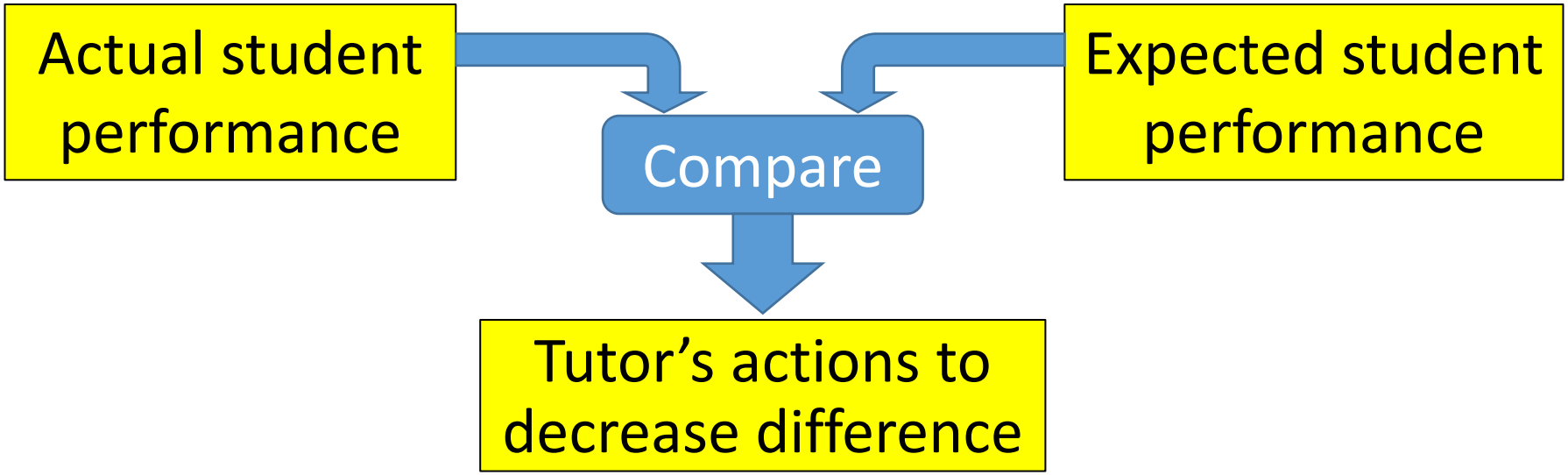
Kurt VanLehn
Arizona State University
Tempe, AZ

Soller et al. (2005) claim edutech innovation is either:

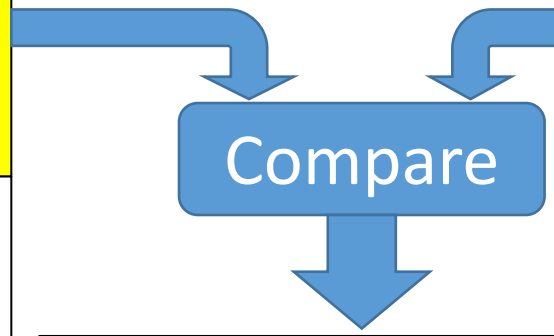
- **Structural:** Changes the lesson plan, content and activities
- **Regulative:** Adds a regulative (i.e., cybernetic; feedback) loop:
  - **Sense** the students' performance
  - **Compare** the students' performance to **Expectations**
  - **Act** to decrease Δ between the students' actual and expected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

- Soller, A., Martinez, A., Jermann, P., & Muehlenbrock, M. (2005). From mirroring to guiding: A review of state of the art technology for supporting collaborative learning. *International Journal of Artificial Intelligence and Education, 15*, 261-290.
- VanLehn, K. (2016). Regulative loops, step loops and task loops. *International Journal of Artificial Intelligence in Education, 26*(1), 107-112

Main components of an ITS Viewed as a regulative loop



To design an ITS, choose at least one from each column



Actual student performance

Sources:

- ❑ Answers to questions
- ❑ Essays
- ❑ Actions in a game
- ❑ Etc.

Tutor's actions to decrease difference

Action types:

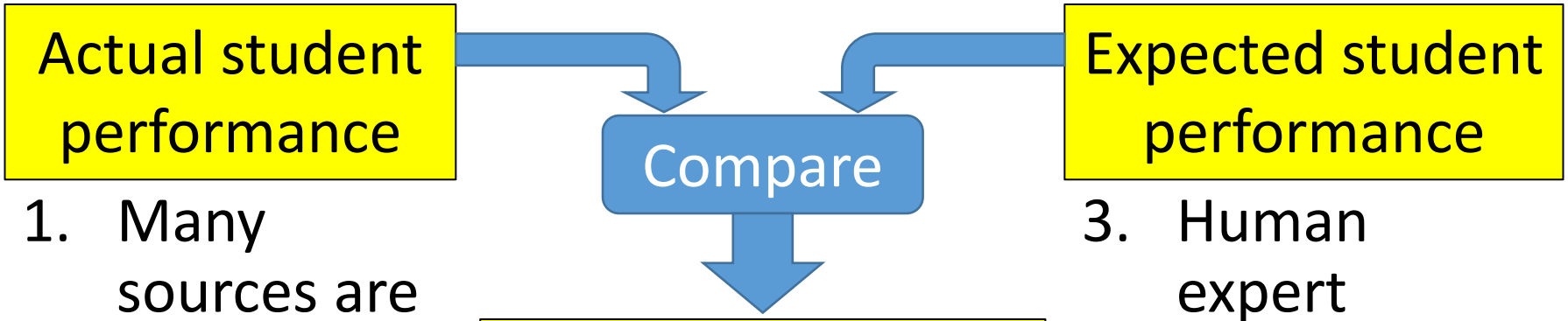
- ❑ Give feedback
- ❑ Choose next task
- ❑ Etc.

Expected student performance

Sources:

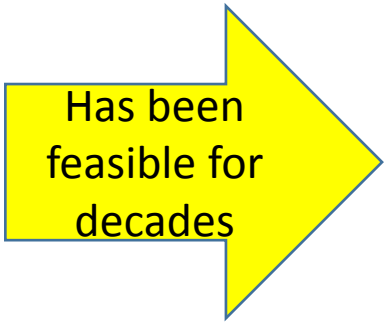
- ❑ Expert authors
- ❑ Algorithms
- ❑ Etc.

I will make 4 main points

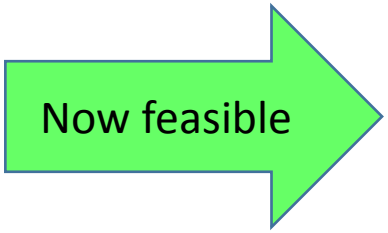


1. Many sources are feasible
2. More frequent data are better, up to a point.
3. Human expert authors are (still) the main source.
4. Only three tutor action types have strong evidence of effectiveness.

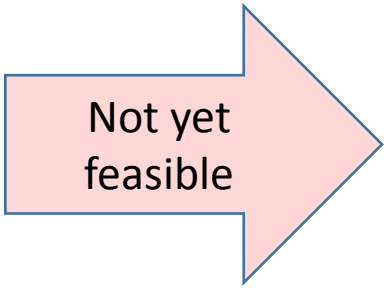
Main sources of student performance data



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

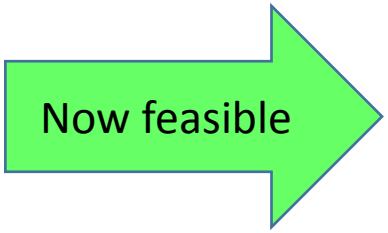


- Step-based
  - Tutor assigns task, then student makes many actions observed by the tutor (steps).



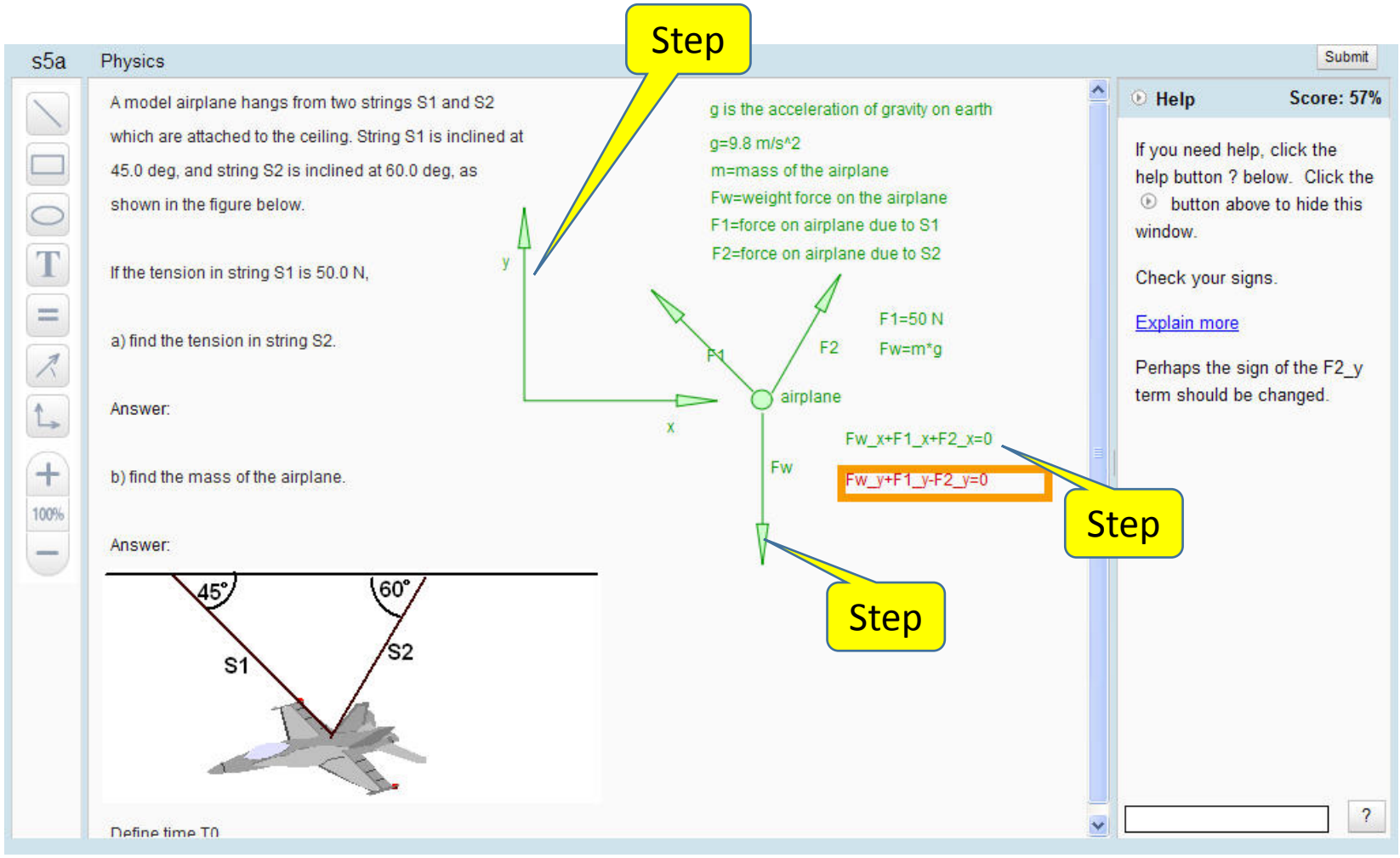
- Spoken student discussions
  - Tutor assigns task, then a small group of students discuss orally.

Next 7 slides are examples of step-based tutors' user interfaces



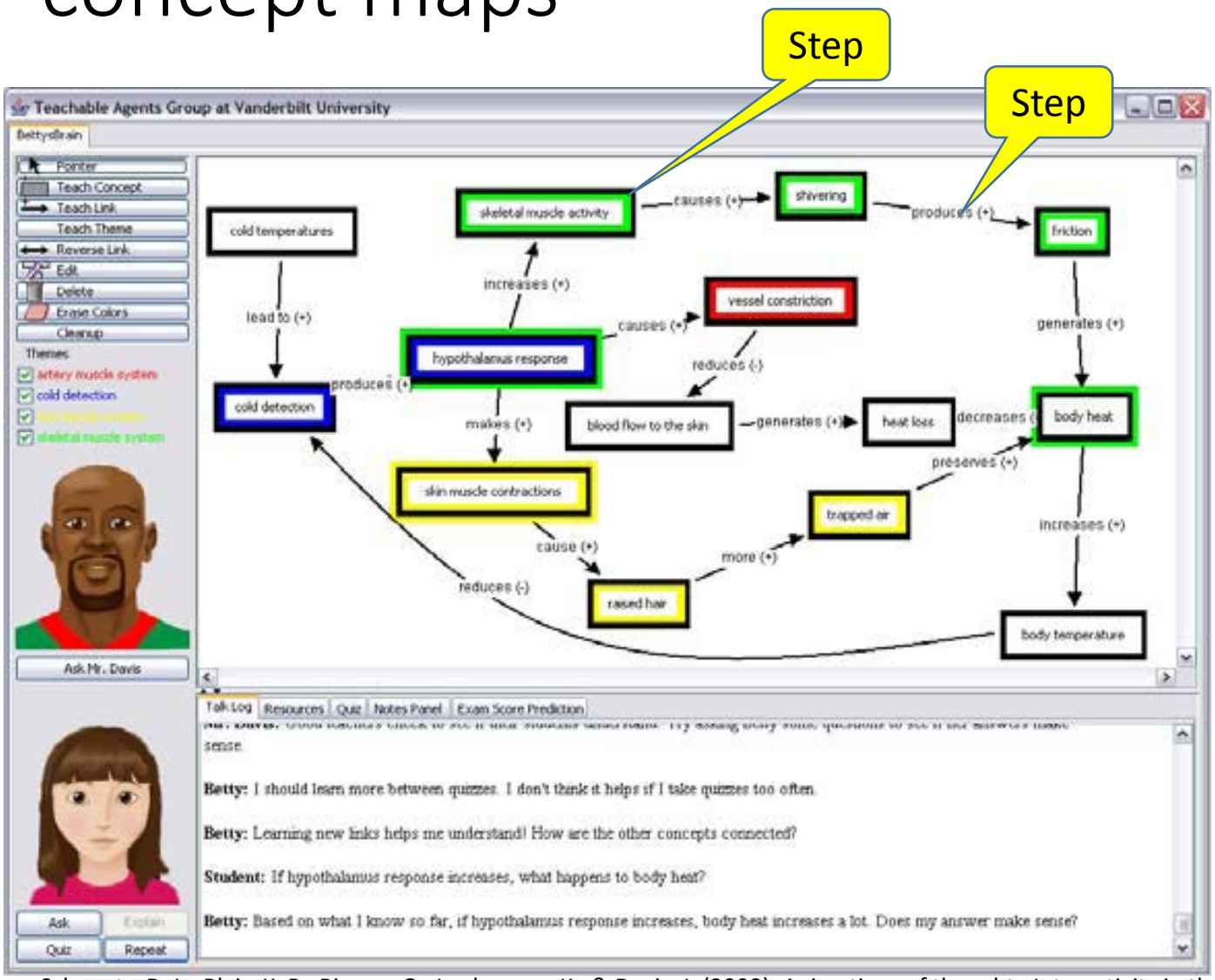
- Step-based
  - Tutor assigns task, then student makes many actions observed by the tutor (steps).

An editor for solving physics problems



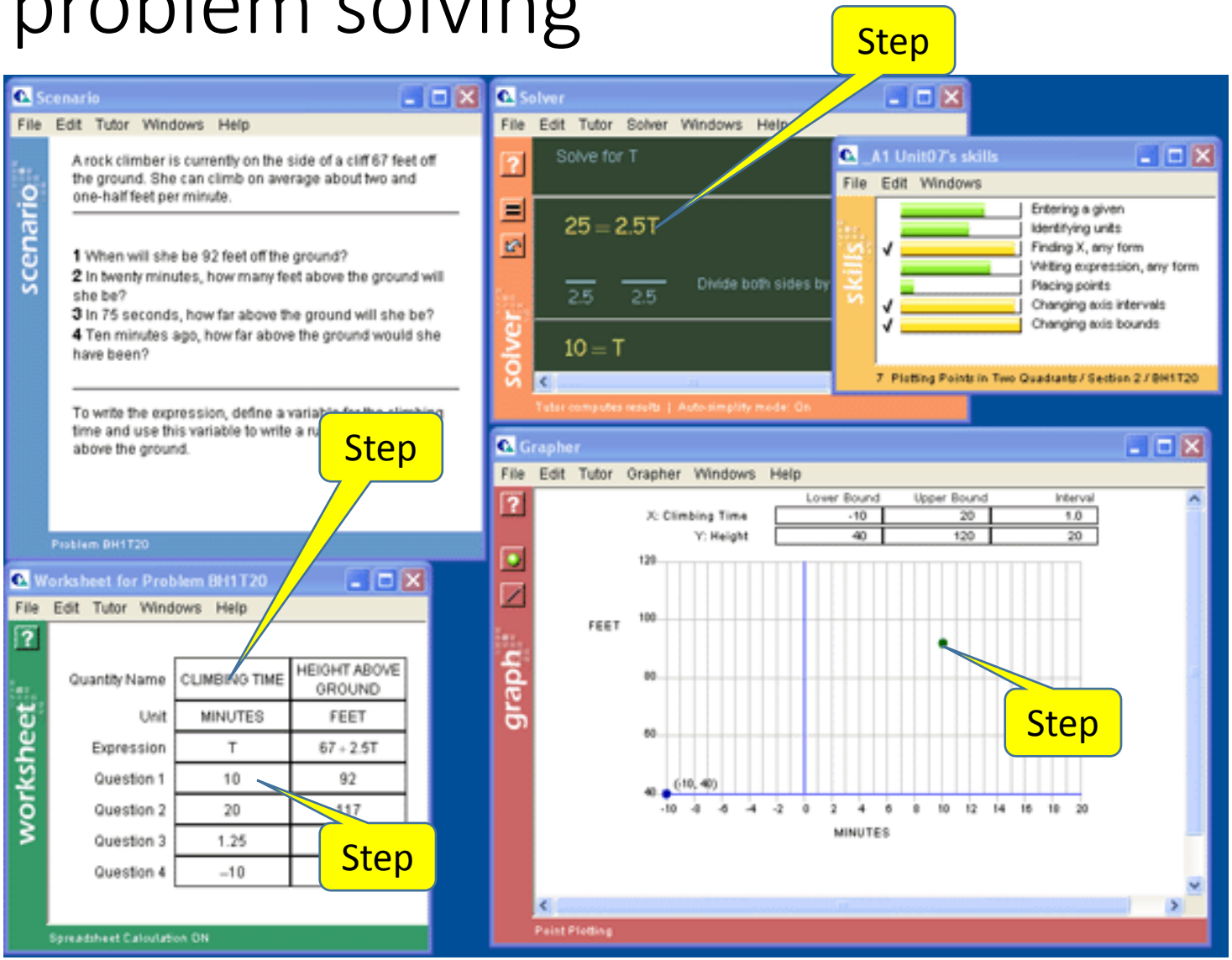
VanLehn, K., Lynch, C., Schultz, K., Shapiro, J. A., Shelby, R. H., Taylor, L., et al. (2005). The Andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence and Education, 15(3), 147-204.*

An editor for constructing concept maps



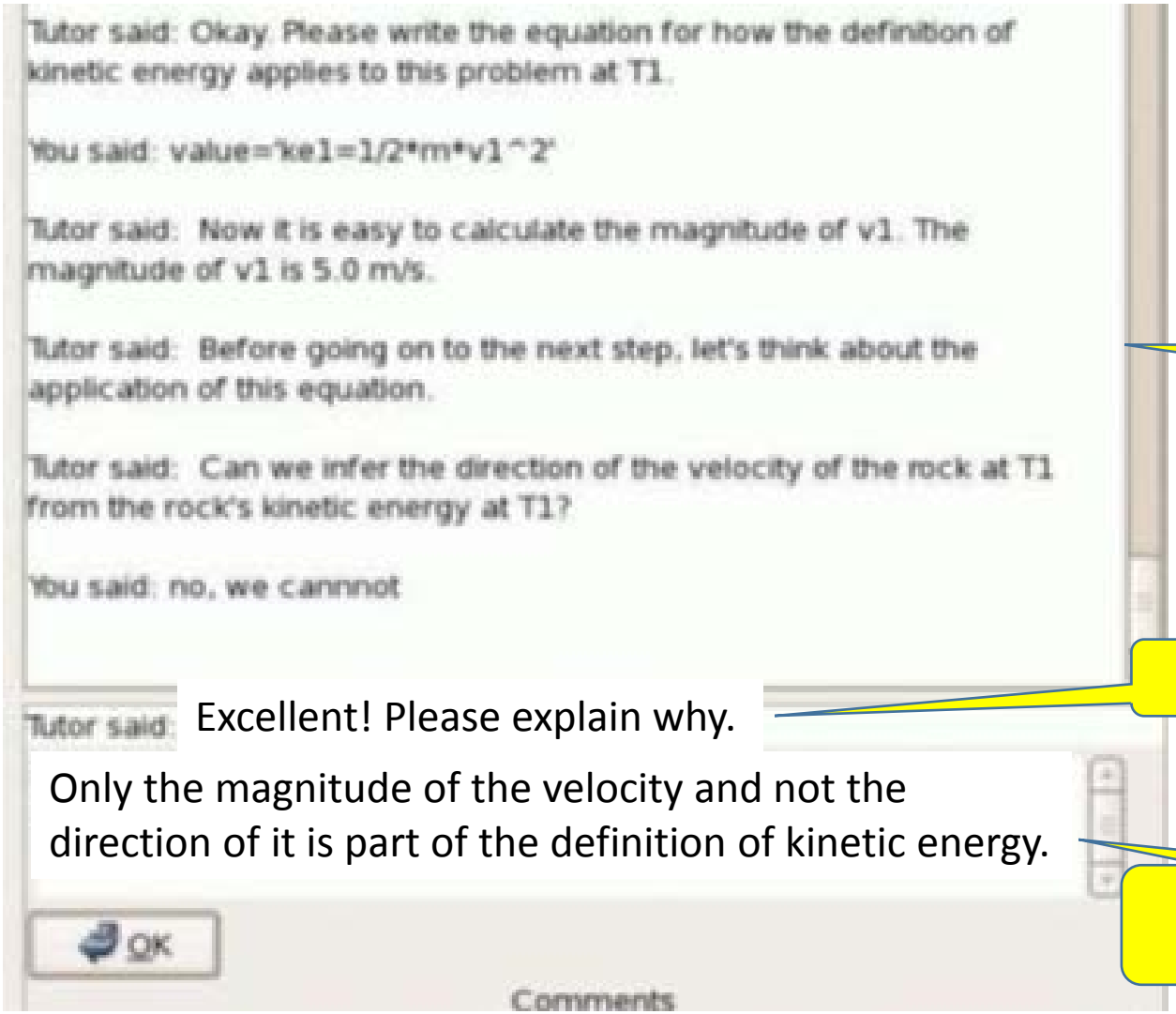
Schwartz, D. L., Blair, K. P., Biswas, G., Leelawong, K., & Davis, J. (2008). Animations of thought: Interactivity in the teachable agent paradigm. In R. Lowe & W. Schnotz (Eds.), *Learning with animations: Research and implications for design. Cambridge, UK: Cambridge University Press.*

An editor for complex math problem solving



www.carnegielearning.com

Tutor-student dialogue



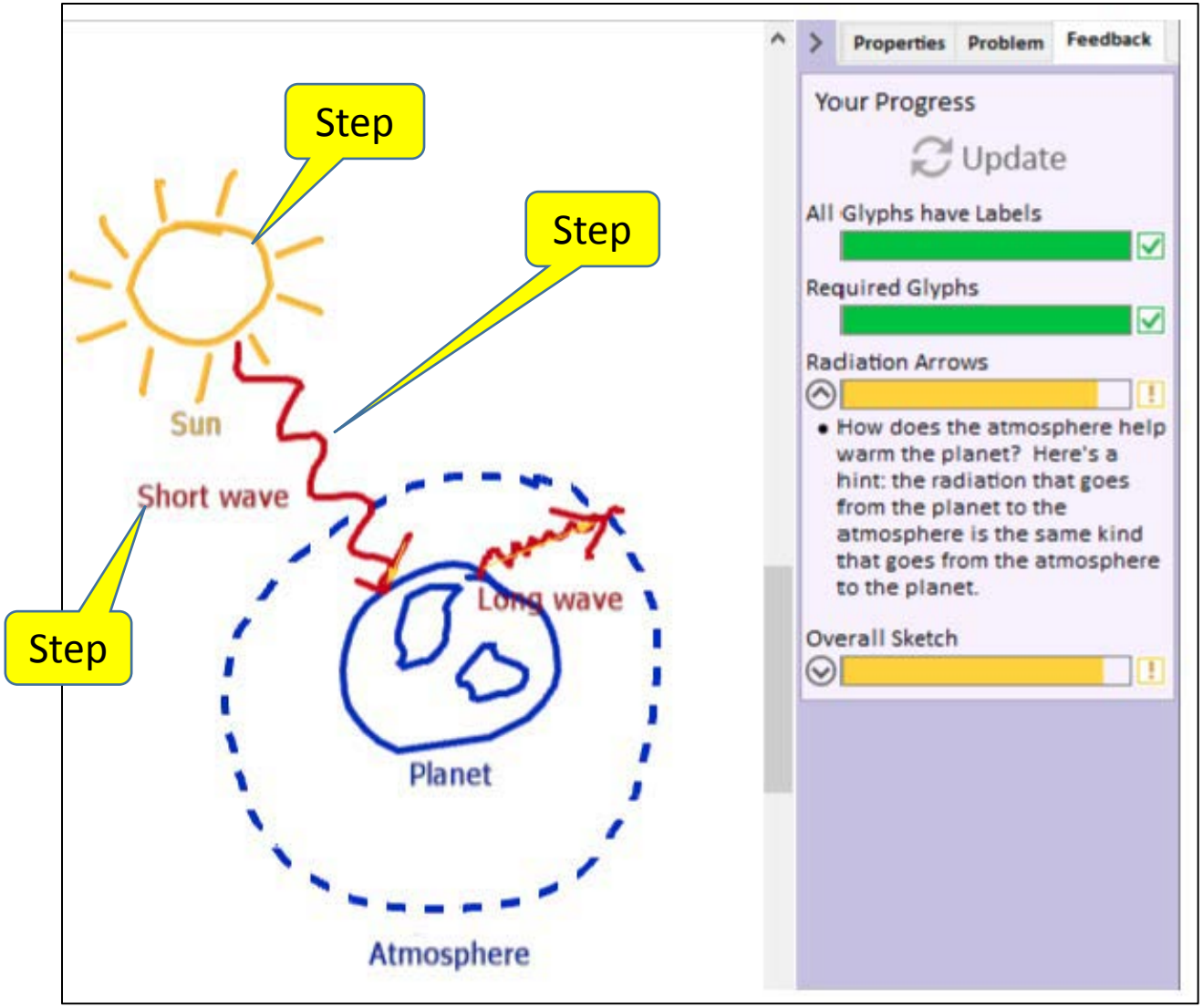
Past dialogue

Tutor's dialog turn

Step:
Student's dialog turn

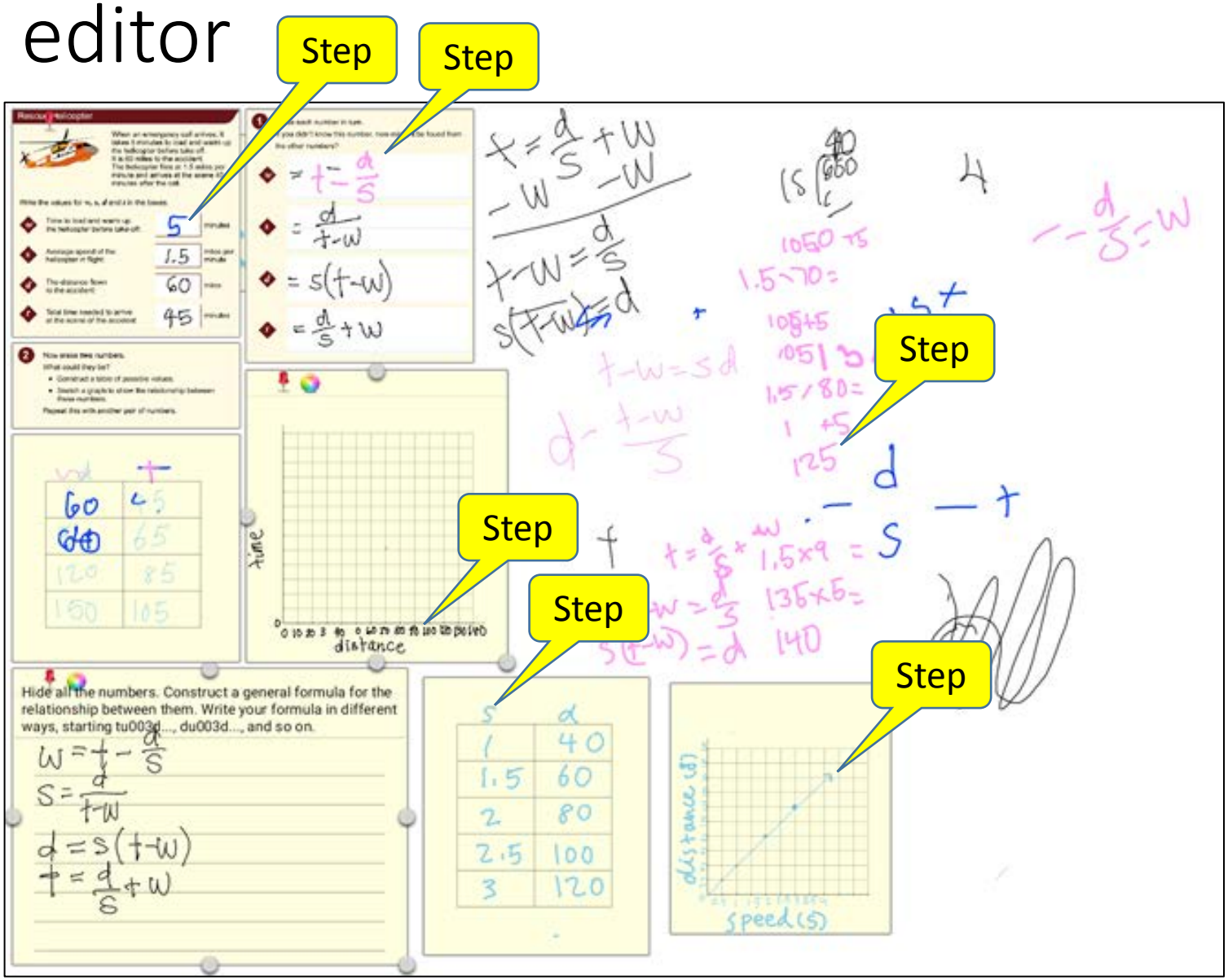
Chi, M., Jordan, P., & VanLehn, K. (2014). When is tutorial dialogue more effective than step-based tutoring? *International Conference on Intelligent Tutoring Systems* (pp. 210-219). Berlin: Springer.

An editor for drawing explanations



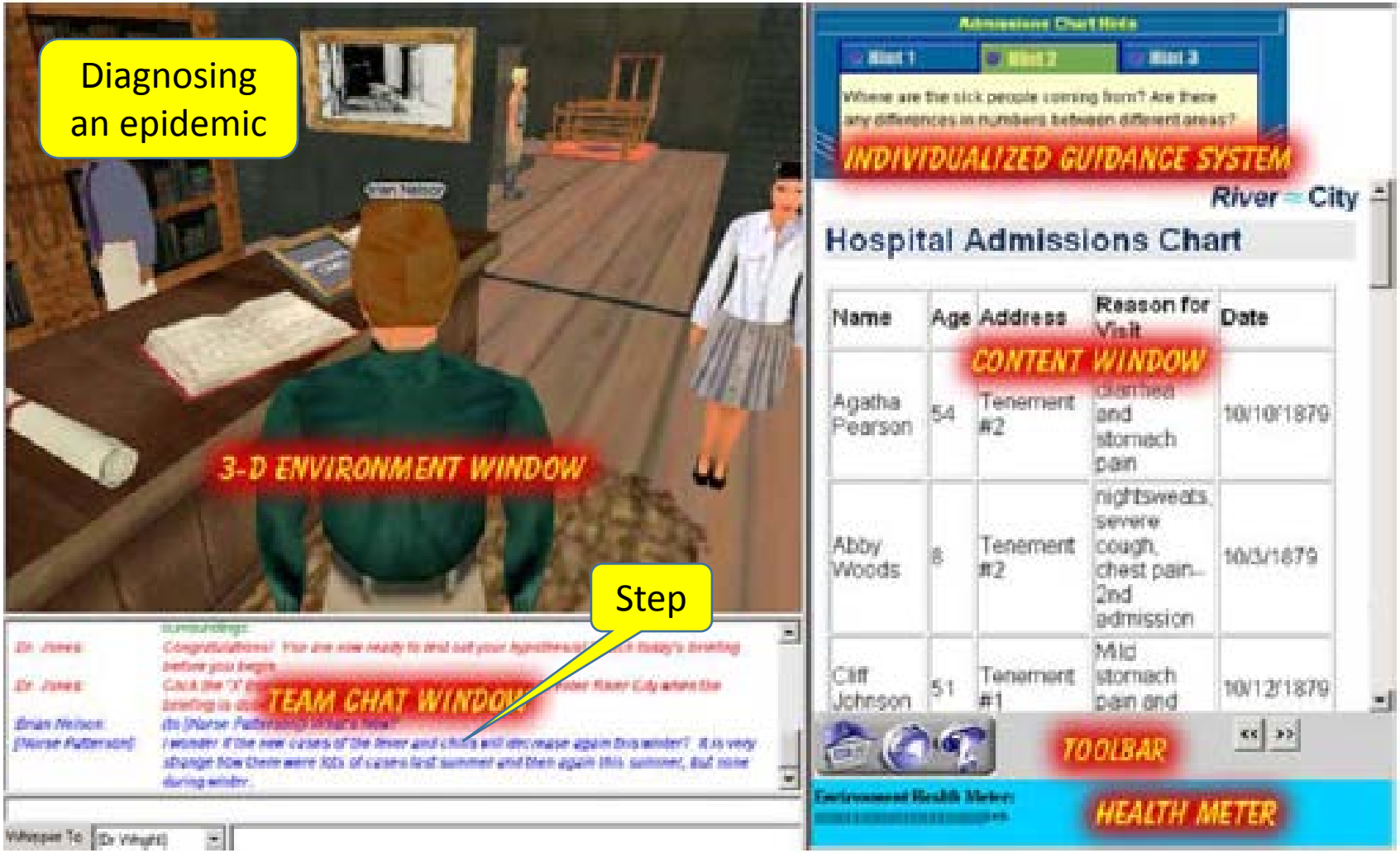
Forbus, K. D. (2016). Sketch understanding for education. In R. Sottilare, A. C. Graesser, X. Hu, A. Olney, B. d. Nye, & A. M. Sinatra (Eds.), Design Recommendations for Intelligent Tutoring Systems: Volume 4 Domain Modeling (pp. 225-235): US Army Research Laboratory.

A general-purpose collaborative editor



VanLehn, K., Burkhardt, H., Cheema, S., Pead, D., Schoenfeld, A. H., & Wetzel, J. (submitted). How can a classroom orchestration system help math teachers improve collaborative, productive struggle?

A multiplayer game



Nelson, B. C. (2007). Exploring the use of individualized, reflective guidance in an educational multi-user virtual environment. Journal of Science Education and Technology, 16(1), 83-97.

How close are tutors to understanding unconstrained speech?

Has been feasible for decades

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

Now feasible

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

Not yet feasible

- 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

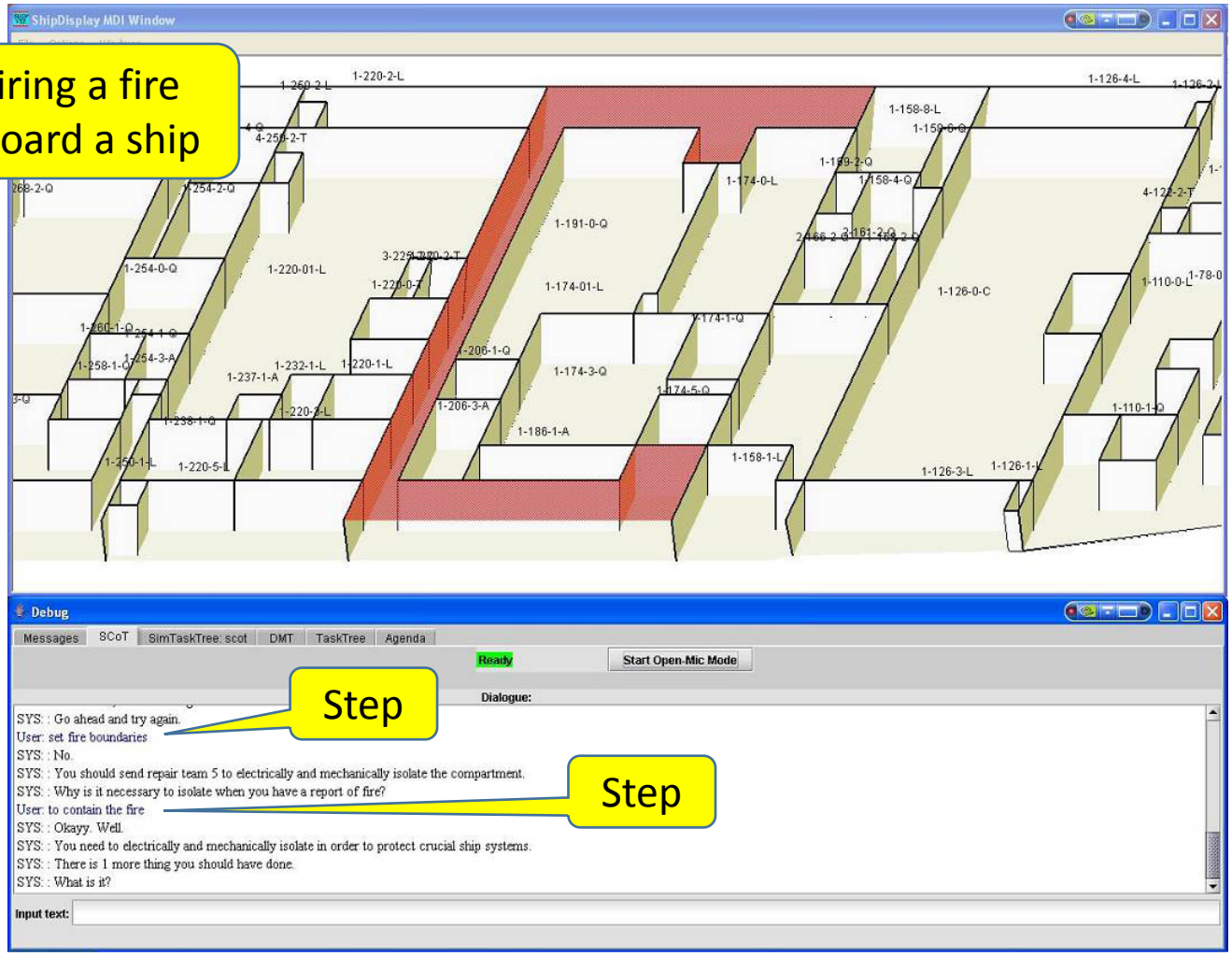
Step

Step

http://www.netc.navy.mil/centers/swos/Simulators.htm

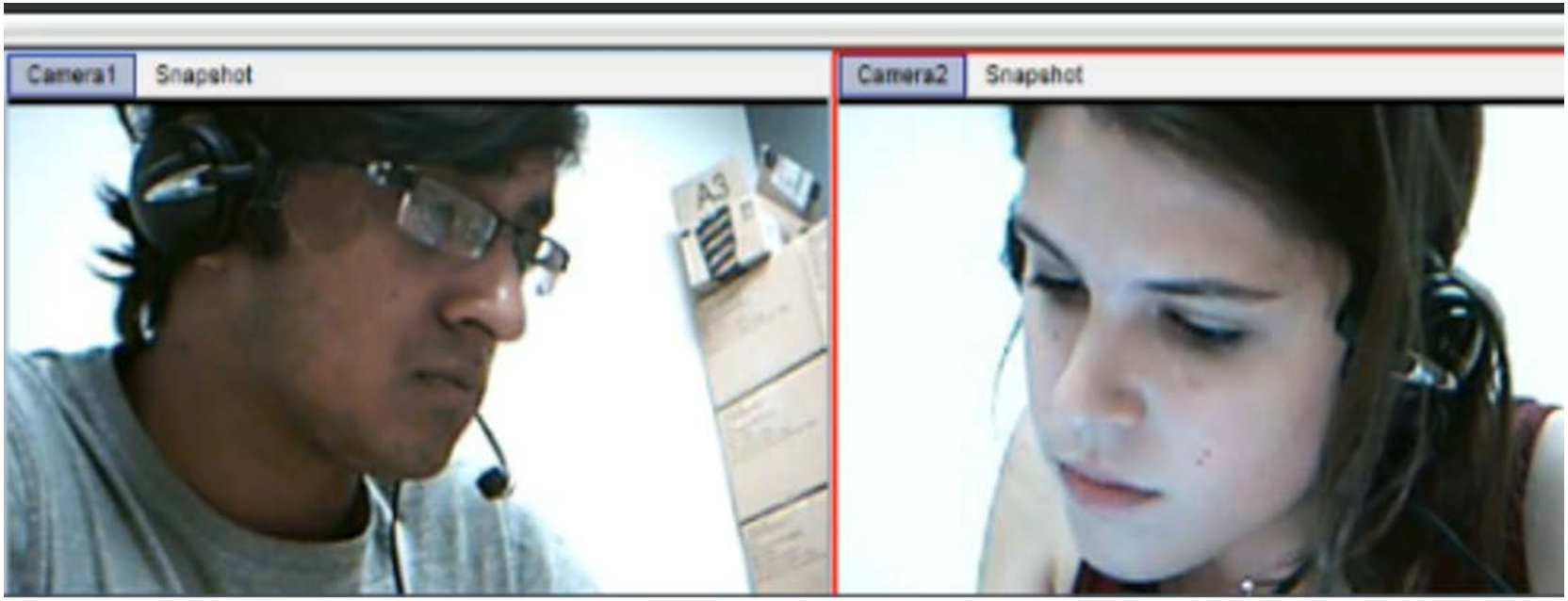
Tutors can understand short answers to their questions

Firing a fire aboard a ship



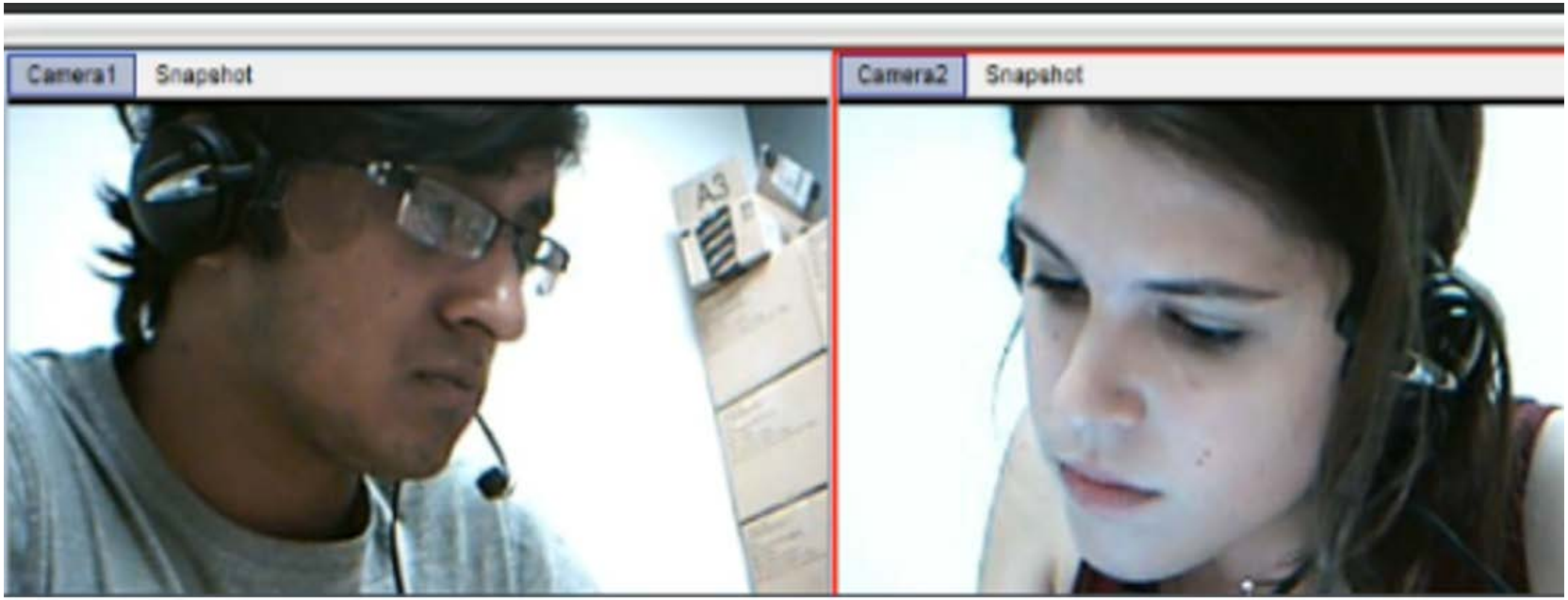
Pon-Barry, H., Schultz, K., Bratt, E. O., Clark, B., & Peters, S. (2006). Responding to student uncertainty in spoken tutorial dialogue systems. International Journal of Artificial Intelligence and Education, 16, 171-194.

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

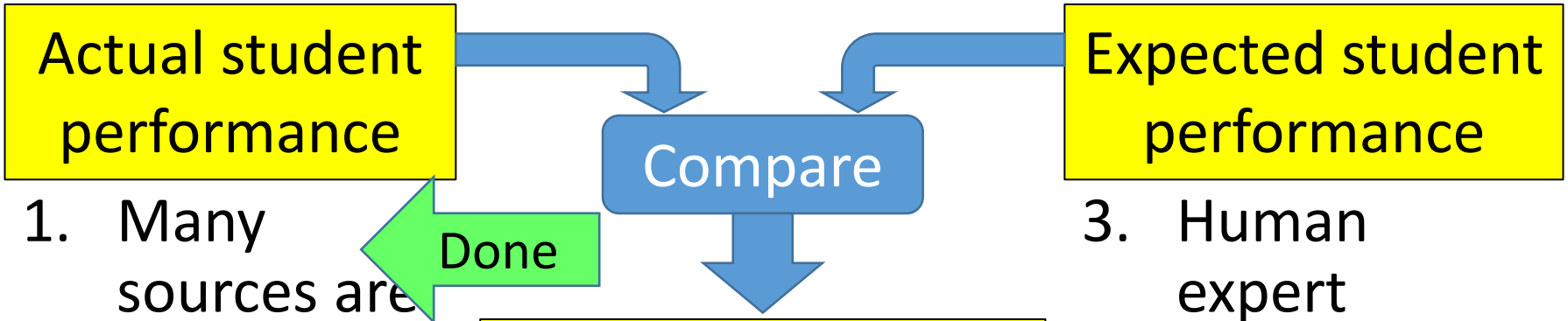


- Viswanathan, S. A., & VanLehn, K. (in press). Using the tablet gestures and speech of pairs of students to classify their collaboration. *IEEE Transactions on Learning Technologies*.
- Forbes-Riley, K., & Litman, D. (2014). Evaluating a spoken dialogue system that detects and adapts to user affective states. Paper presented at the *SIGDial: 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Philadelphia, PA.

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.

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

Tutoring type	vs. other tutoring type	Num. of effects	Mean effect
Answer-based	no tutoring	165	0.31
Step-based		28	0.76
Human		10	0.79

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

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems and other tutoring systems. *Educational Psychologist, 46(4), 197-221*

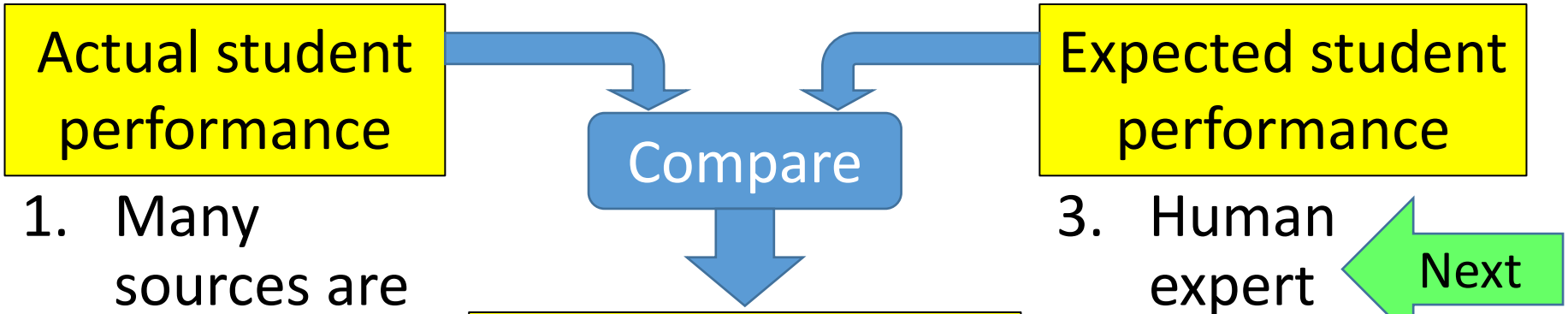
More frequent tutor-student interactions foster more learning, up to a point

Tutoring type	vs. other tutoring type	Num. of effects	Mean effect
Answer-based	no tutoring	165	0.31
Step-based		28	0.76
Human		10	0.79
Step-based	answer-based	2	0.40
Human		1	-0.04
Human	step-based	10	0.21

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

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems and other tutoring systems. *Educational Psychologist, 46(4), 197-221*

Main points: Transition slide



1. Many sources are feasible

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

Done

4. Only three tutor action types have strong evidence of effectiveness.

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

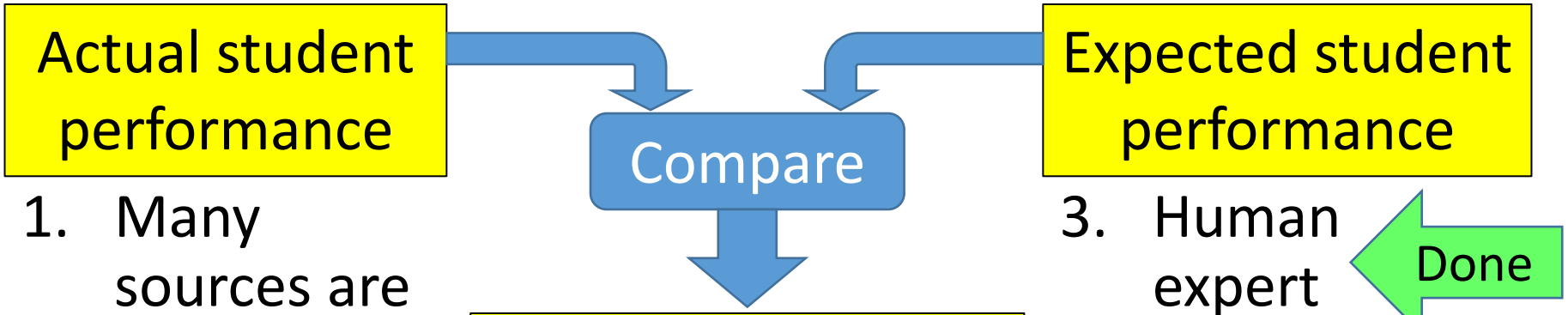
Next

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.

Done

4. Only three tutor action types have strong evidence of effectiveness.

Next

Common activities in classes.

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

Next

ITS are feasible

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

Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., . . . Thissen, D. (2000). Computerized adaptive testing: A primer (second ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Corbett, A., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction, 4*, 253-278.

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

Kulik, C., Kulik, J., & Bangert-Drowns, R. (1990). Effectiveness of mastery learning programs: A meta-analysis. *Review of Educational Research, 60*(2), 265-306.

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

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research, 86*(1), 42-78.

Adaptive task selection: Weak evidence or effect

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

Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. A. (2008). Learning styles: Concepts and evidence. *Psychological Science in the Public Interest, 9*(3), 105-119.

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

Karich, A. C., Burns, M. K., & Maki, K. E. (2014). Updated meta-analysis of learner control within educational technology. *Review of Educational Research, 84*(3), 392-410.

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

Sense, F., Behren, F., Meijer, R.B. & van Rijn, H. (2016) An individual's rate of forgetting is stable over time but differs across materials. *Topics in Cognitive Science*, 8(1), 305-321.

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 unmastered topics
  - Just one study?

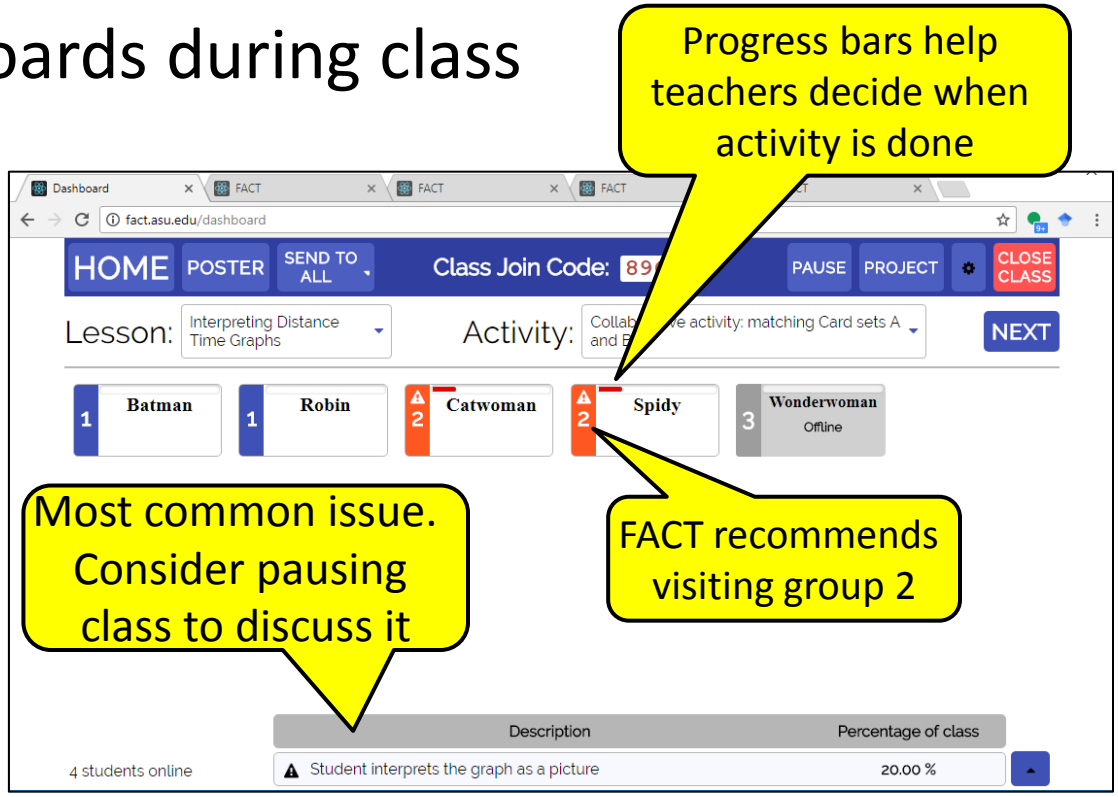
Koedinger, K. R., Pavlik, P. I., Stamper, J., Nixon, T., & Ritter, S. (2011). *Avoiding problem selection thrashing with conjunctive knowledge tracing*. Paper presented at the International Conference on Educational Data Mining, Eindhoven, NL.

ITS impact on small group work: Weak evidence

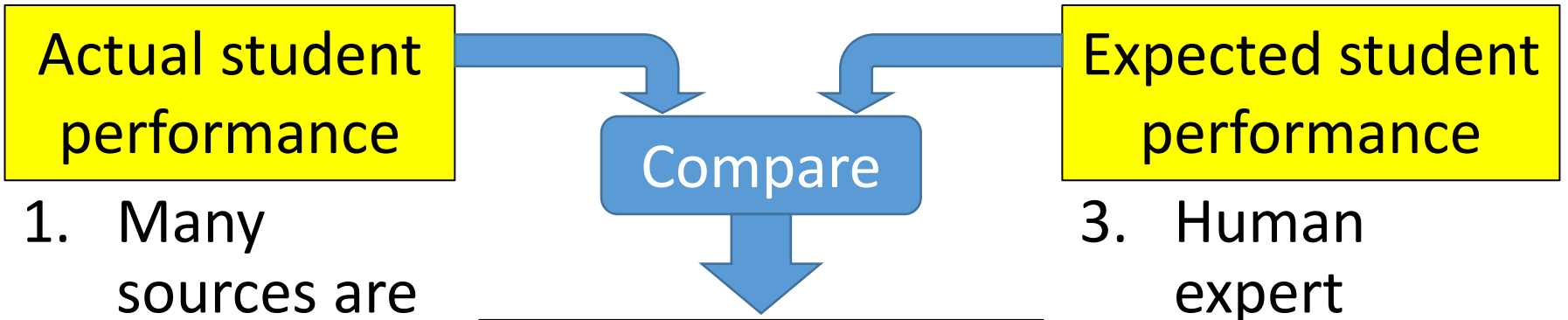
- Feedback and hints
  - Most studies focus on increasing collaboration
  - Few studies measure learning
- Selecting group members
  - Few studies measure learning

ITS impact on teachers: Weak evidence

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



Summary



1. Many sources are feasible
2. More frequent data are better, up to a point.

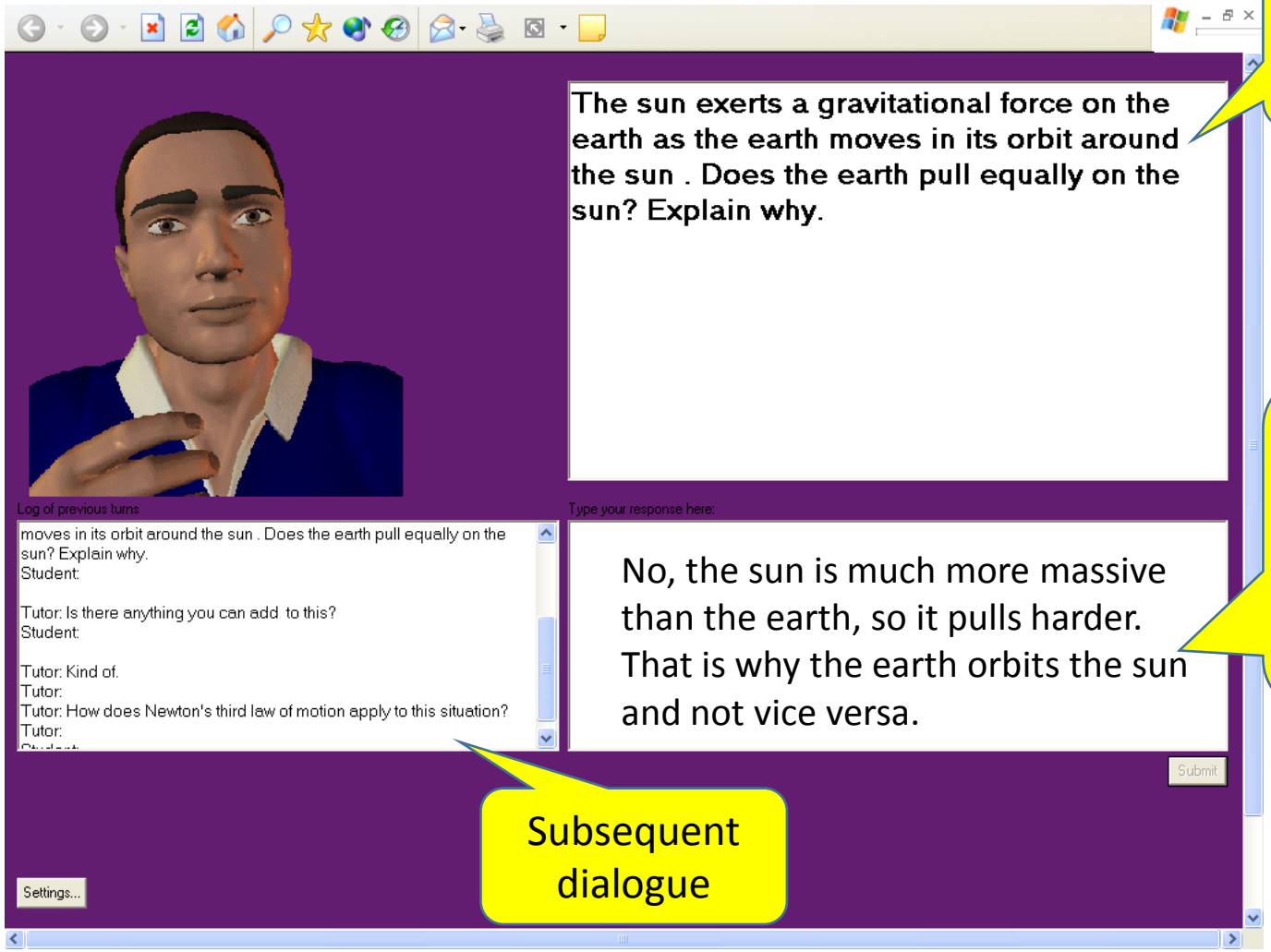
Tutor's actions to decrease difference

- 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

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

Thanks!
Questions?

A tutor-student dialogue that starts with an essay question



Graesser, A. C., D'Mello, S. K., Hu, X., Cai, Z., Olney, A., & Morgan, B. (2012). AutoTutor. In P. McCarthy & C. Boonthum-Denecke (Eds.), Applied natural language processing: Identification, investigation and resolution (pp. 169-187). Hershey, PA: IGI Global.