Now that ITS are as effective as human tutors, how can they become even better?

Kurt VanLehn

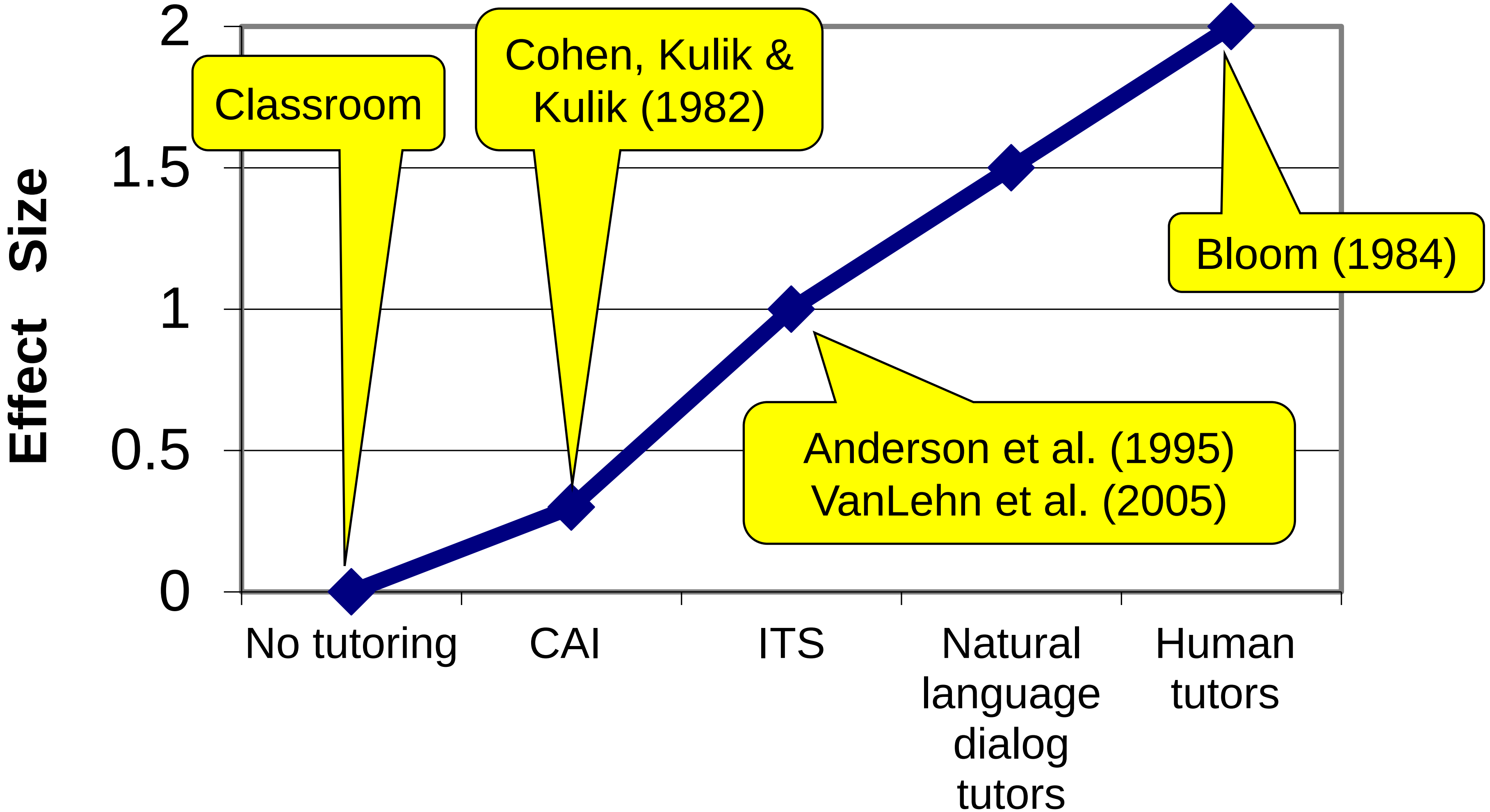
School of Computing, Informatics and Decision Systems Engineering

Arizona State University

Outline

- The interaction granularity hypothesis
  - The smaller the grain size of interaction, the more effective the tutoring
  - Grain size: Human < ITS < CAI < no tutoring
  - Effectiveness? Human > ITS > CAI > no tutoring
- Evidence *against* the hypothesis
  - Effectiveness! Human = ITS > CAI > no tutoring
  - The interaction plateau hypothesis
- How to achieve ITS > Human effectiveness

A widely held belief: Human tutors are much more effective than computer tutors



Why are human tutors so effective? Summary of ~20 studies:

Weak evidence

- Detailed diagnosis
- Personalized task selection
- Sophisticated tutoring strategies
- Learner control
- Broader knowledge
- Motivation

Strong evidence

- Hints
  - – push reasoning along
- Feedback
  - – catch errors quickly

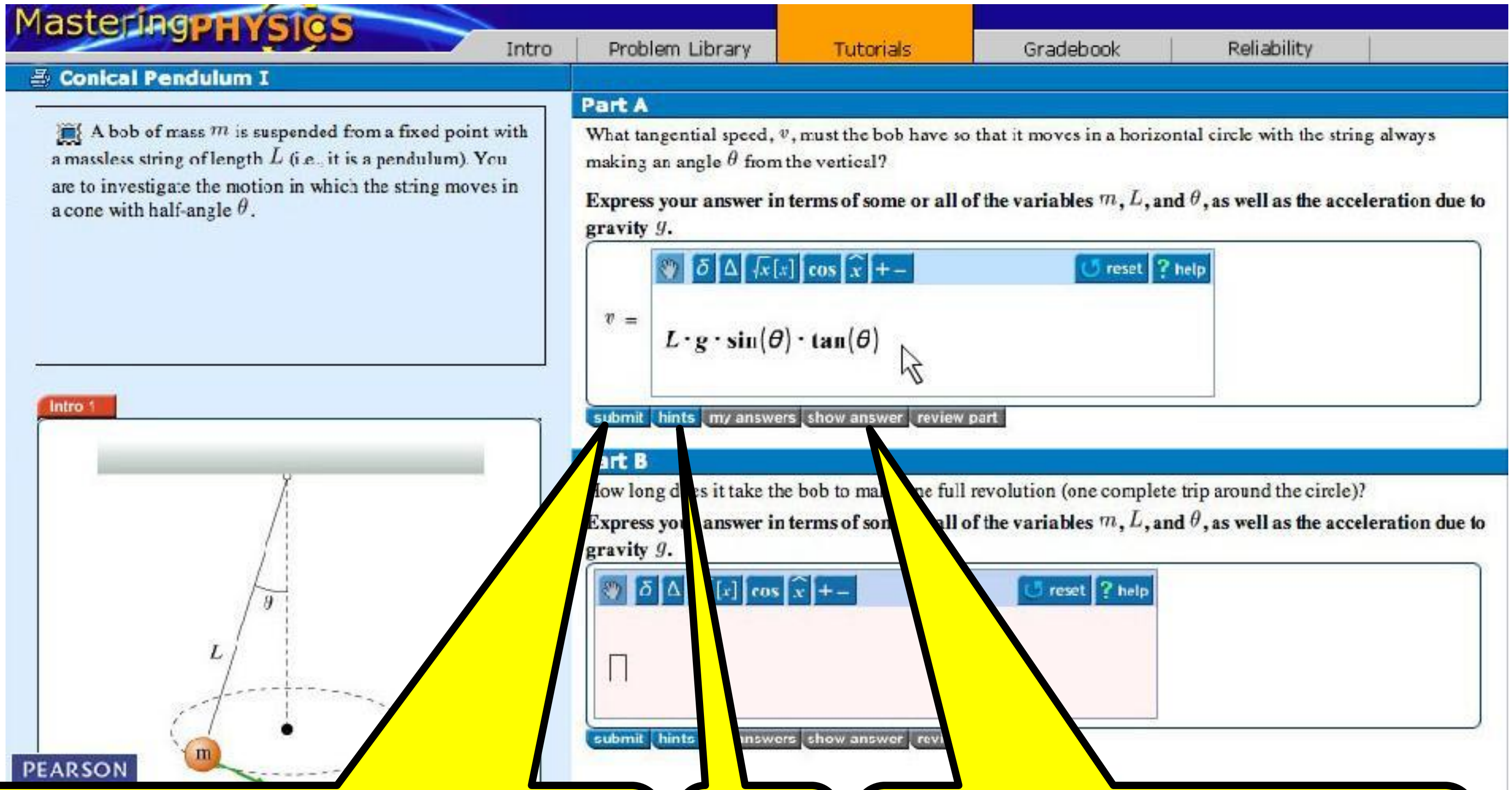
Both human and computer tutors do hinting and feedback

- So why are human tutors more effective?

Both human and computer tutors do scaffolding and feedback

- So why are human tutors more effective?
- Interaction granularity hypothesis:
  - Because the **granularity of the interaction** for human tutors is smaller than for computer tutors, human tutors are more effective.
- Granularity of the interaction:
  - CAI: **Answer**
  - ITS with WIMP (windows, icon, menu, pointing) interface: **Step**
  - ITS with natural language dialogue interface: **Substep**
  - Human tutor: **Arbitrarily fine-grained**

Computer-aided instruction (CAI) ➔ Answer-based tutoring

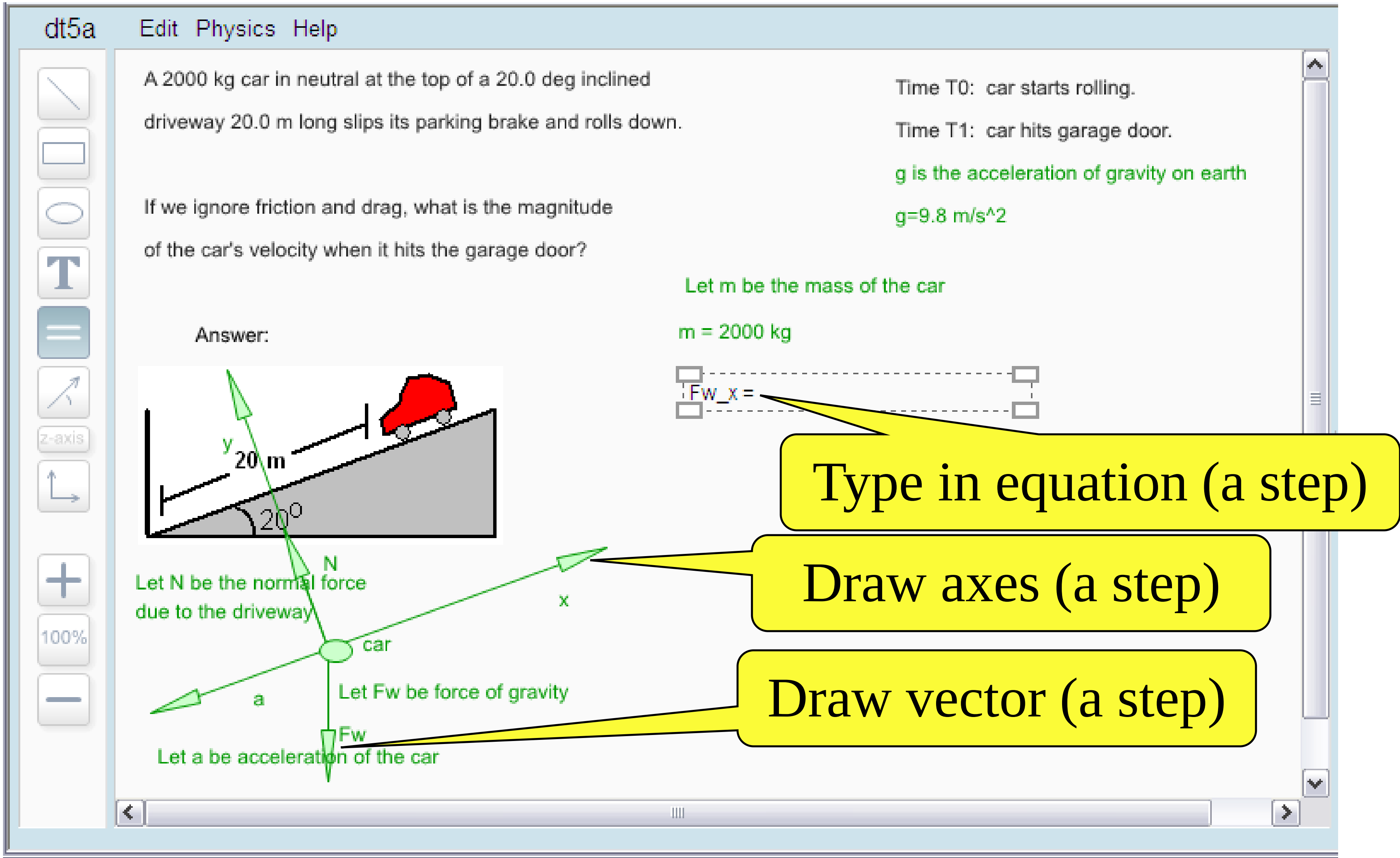


Solve on paper, enter **ANSWER** & get feedback

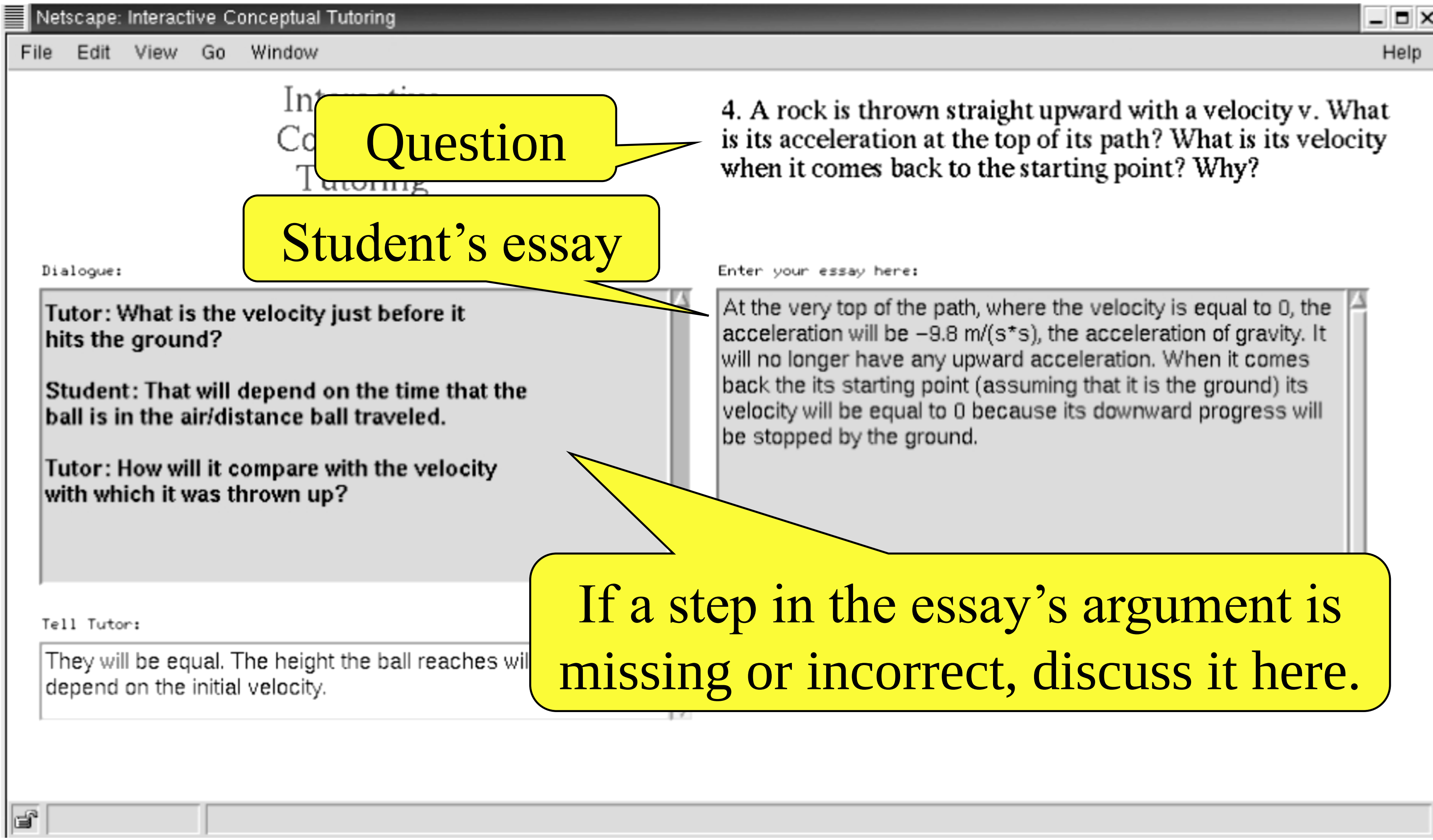
Hints

"Bottom out" hint i.e., correct answer

An ITS (Andes) with WIMP interface ➔ Step-based tutoring



Natural language (NL) dialogue tutoring ➔ substep based tutoring



Human tutoring
➔ arbitrarily fine-grained interaction

T: In a basket of apples, 40% of the apples have worms, and 25% are rotted inside. Assuming independence, what is the chance that a randomly selected apple will have a worm but no rot?

S: I'm stuck

T: Our event is an apple with a worm and no rot, right? — Hint

S: Right.

T: Can that event be decomposed into an AND or an OR or something of two events? — Hint

S: Yes. It's having a worm AND having no rot.

T: Excellent! Can you write that in symbols, like P(...)? — Hint

S: P(worm & rot).

T: Almost. Check your "rot" — Negative feedback

S: P(worm & ~rot)

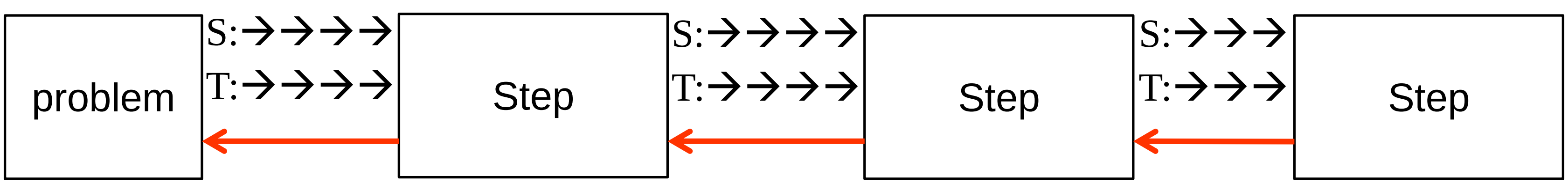
T: Good. Do you know a rule that matches that?

S: P(A&B) = P(A) * P(B)

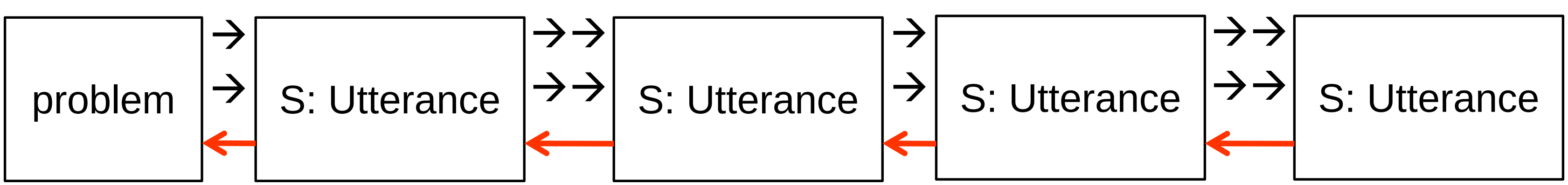
Granularity of tutoring ≈ number of inferences (→) between interactions

- Answer-based tutoring (CAI)

- Step-based tutoring (ITS with WIMP)



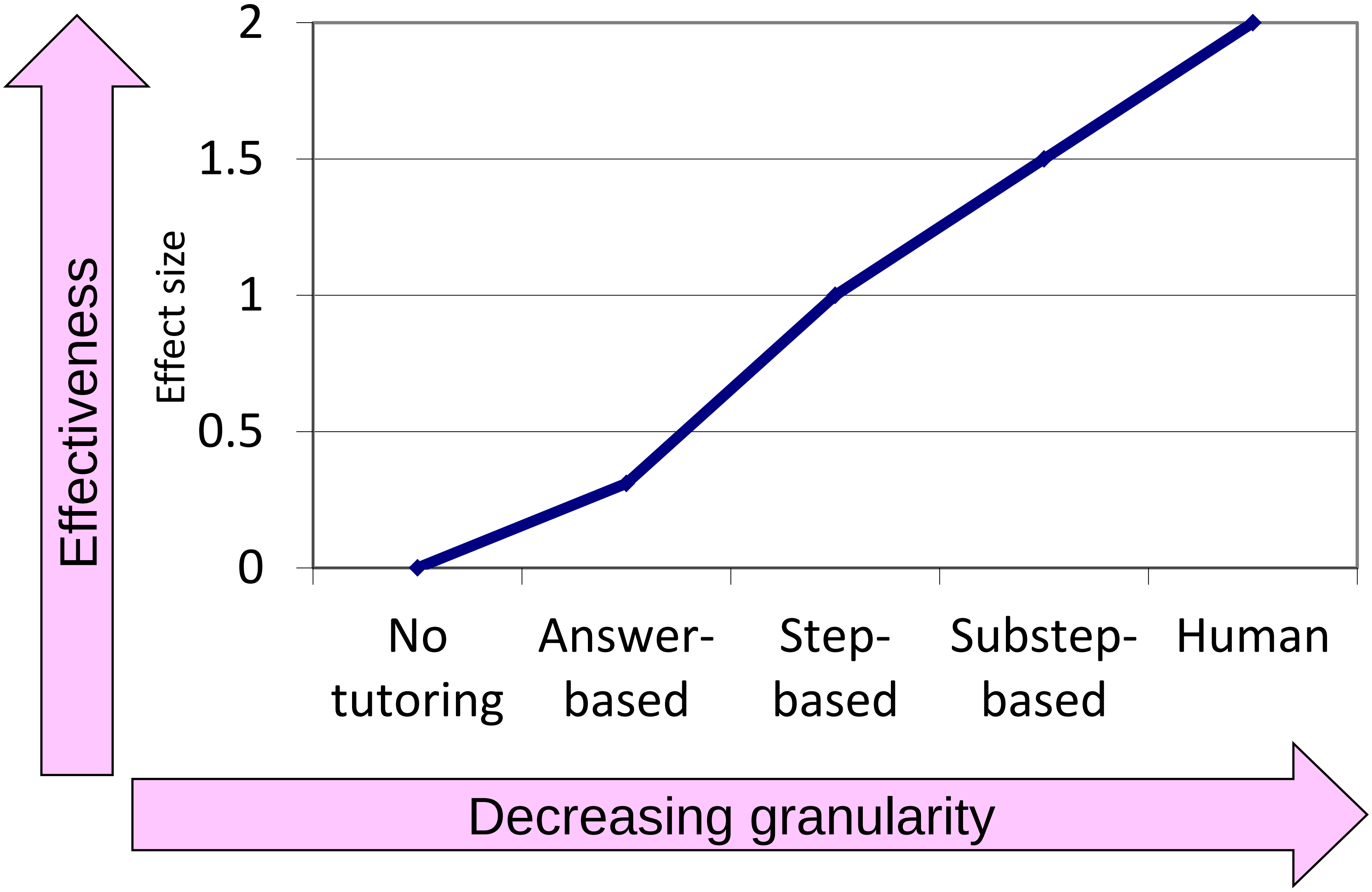
- Human tutoring



The interaction granularity hypothesis

- The smaller the grain size, the more effective the tutoring
- Large grain-size (e.g. Answer-based tutoring)
  - long chain of inferences between allowed interactions
  - negative feedback → where in long chain was the mistake?
  - hinting launches too few inferences to reach answer
- Small grain-size (e.g., Human tutoring)
  - short chain of inferences between allowed interactions
  - negative feedback → debugging short chain is easy
  - hinting launches enough inferences to reach next allowed interaction point

Now the common belief makes sense



Outline

- The interaction granularity hypothesis
  - The smaller the grain size of interaction, the more effective the tutoring
  - Grain size: Human < ITS < CAI < no tutoring
  - Effectiveness? Human > ITS > CAI > no tutoring
- Evidence *against* the hypothesis — Next
  - Effectiveness! Human = ITS > CAI > no tutoring
  - The interaction plateau hypothesis
- How to achieve ITS > Human effectiveness

Andes-Atlas: A substep-based tutoring system

- When Andes detects a conceptual error, it teaches the concept with text
  - About a paragraph (300 words)
- Andes-Atlas conducts dialogue instead
  - called a "knowledge construction dialogue"
  - all text, no audio

Dialogue & text have same content

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

- 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

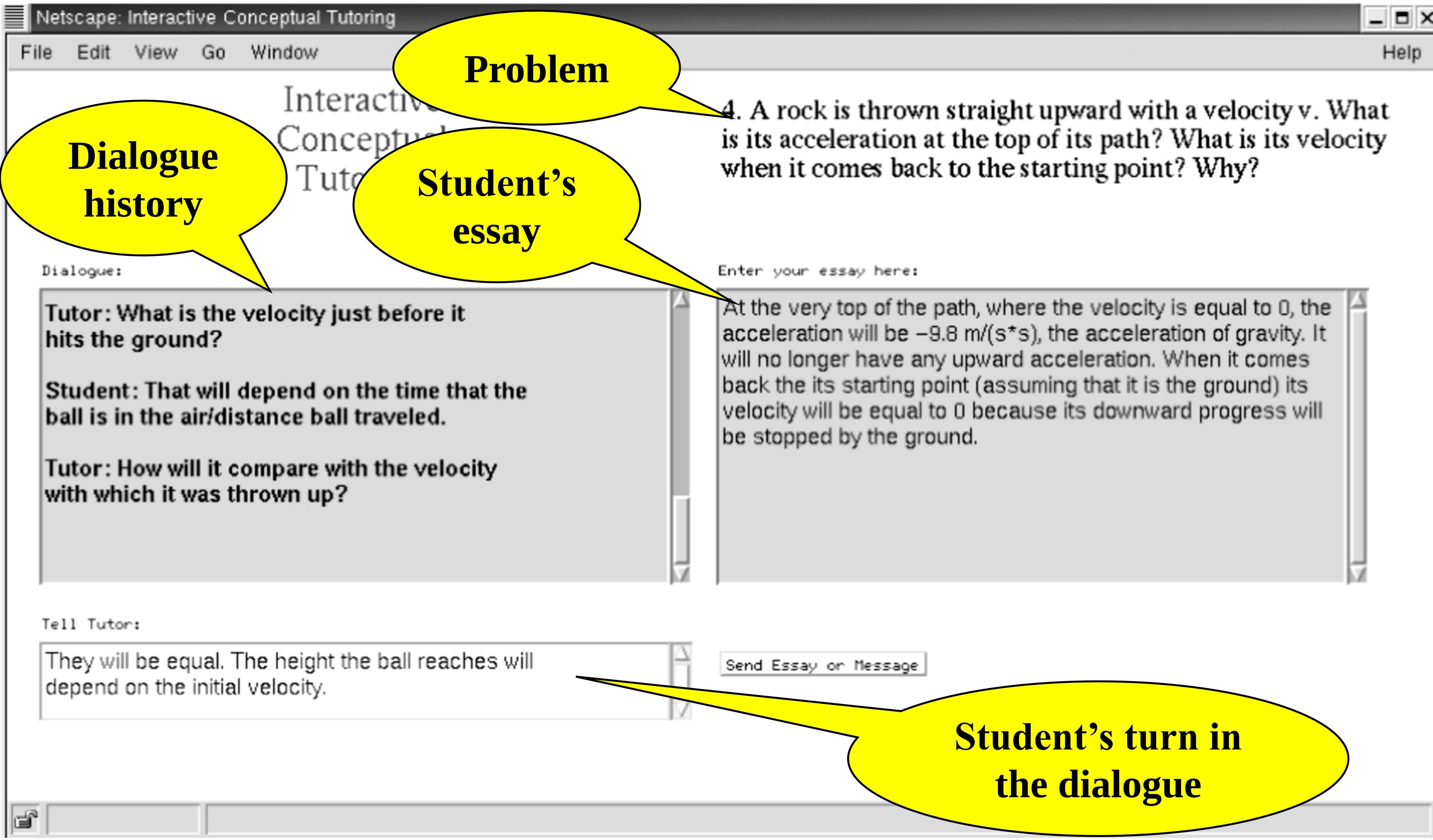
Evidence against the interaction granularity hypothesis: Outline

- √ Andes-Atlas
- Why2 — Next
- Other studies
- Meta-analysis

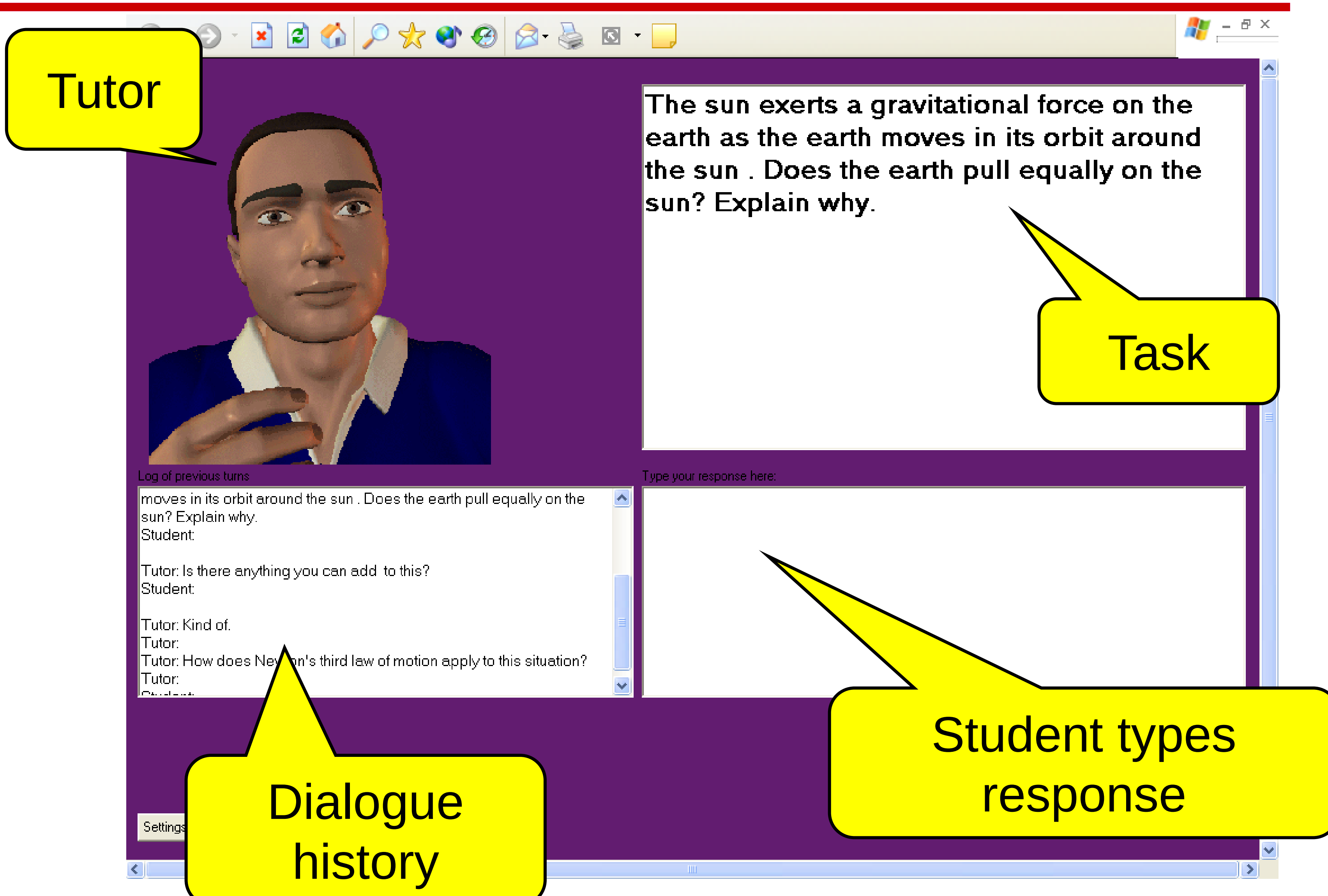
The WHY2 studies

- 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

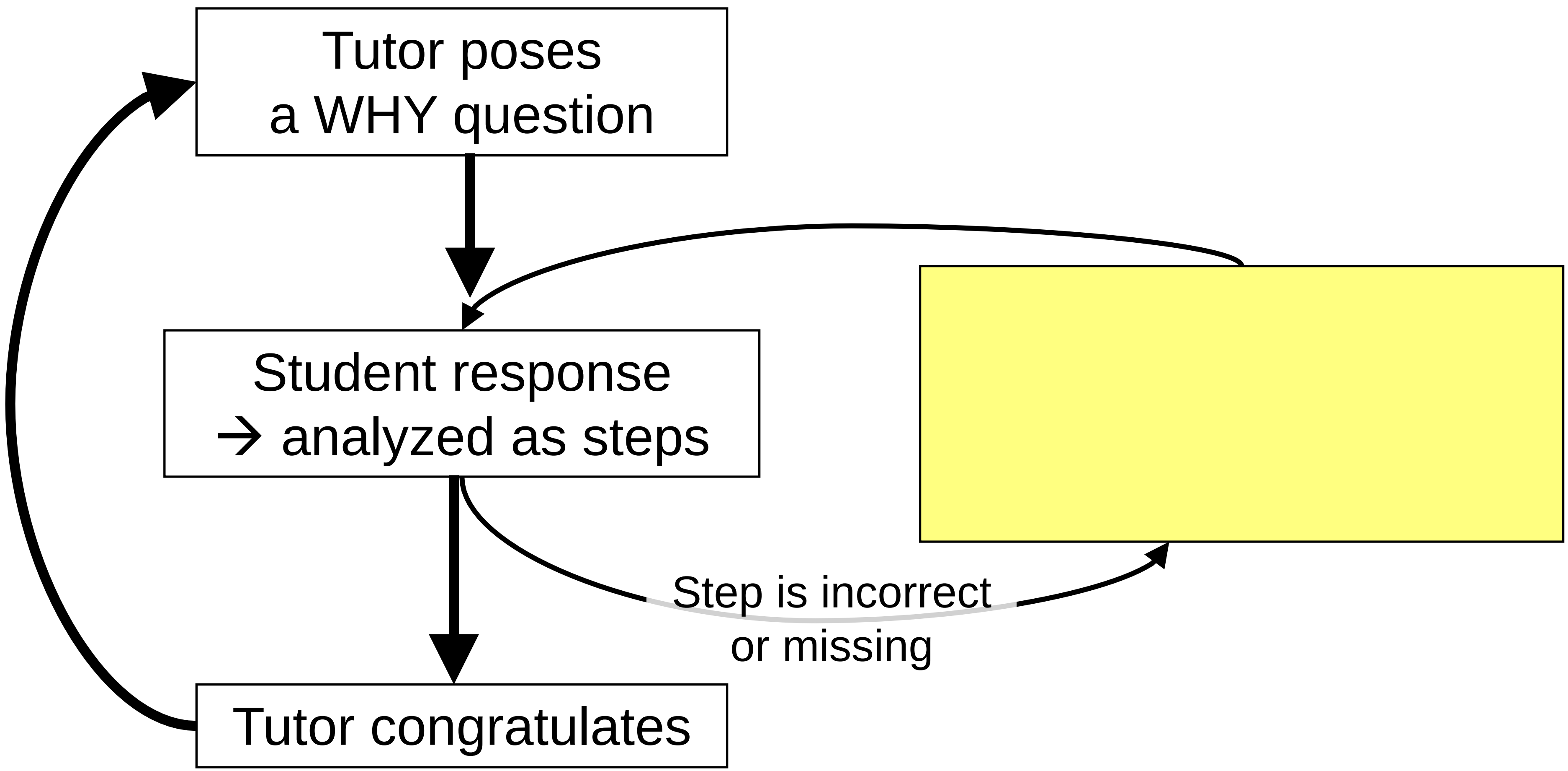
User interface for human tutoring and Why2-Atlas



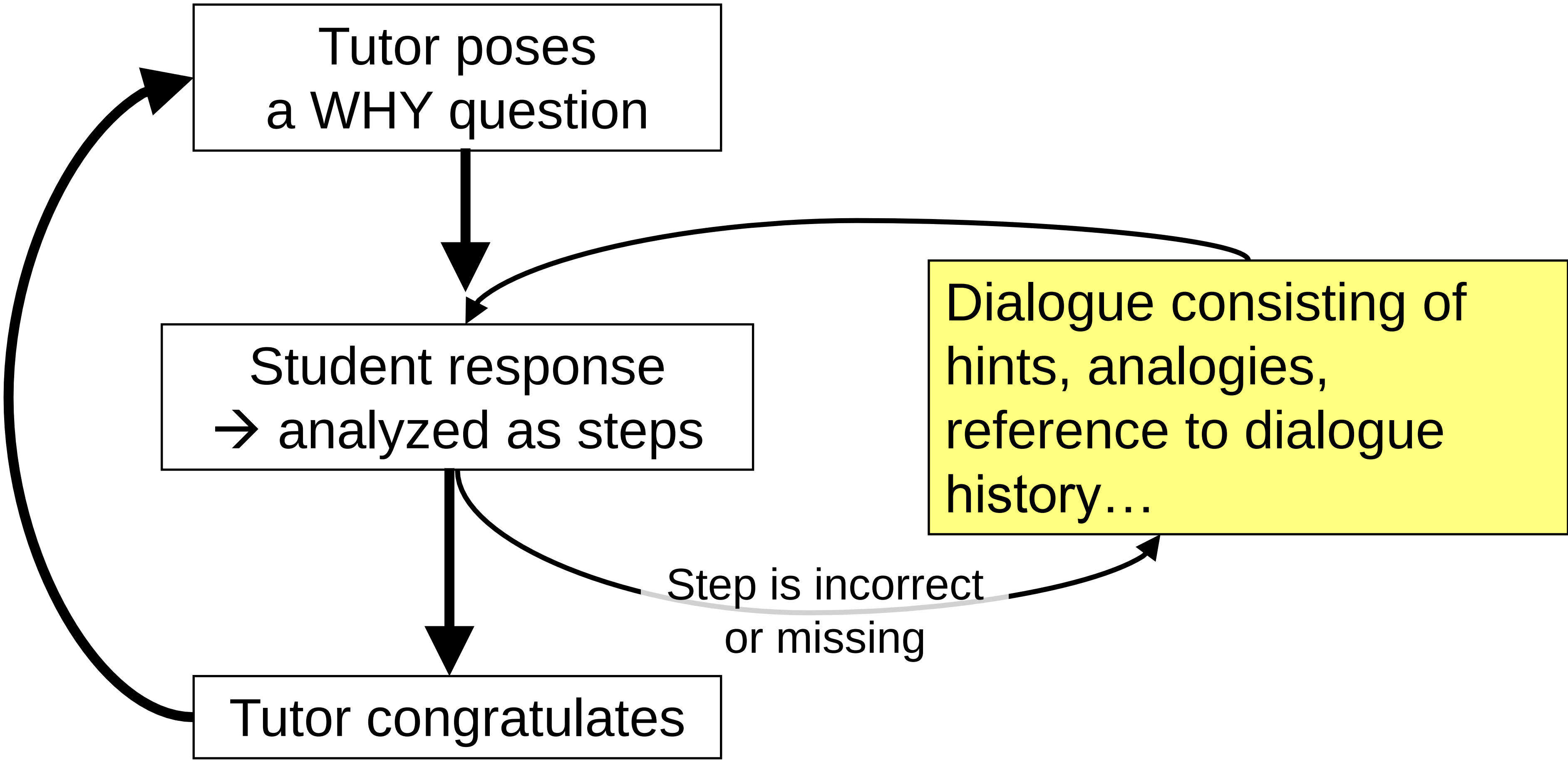
Why2-AutoTutor user interface



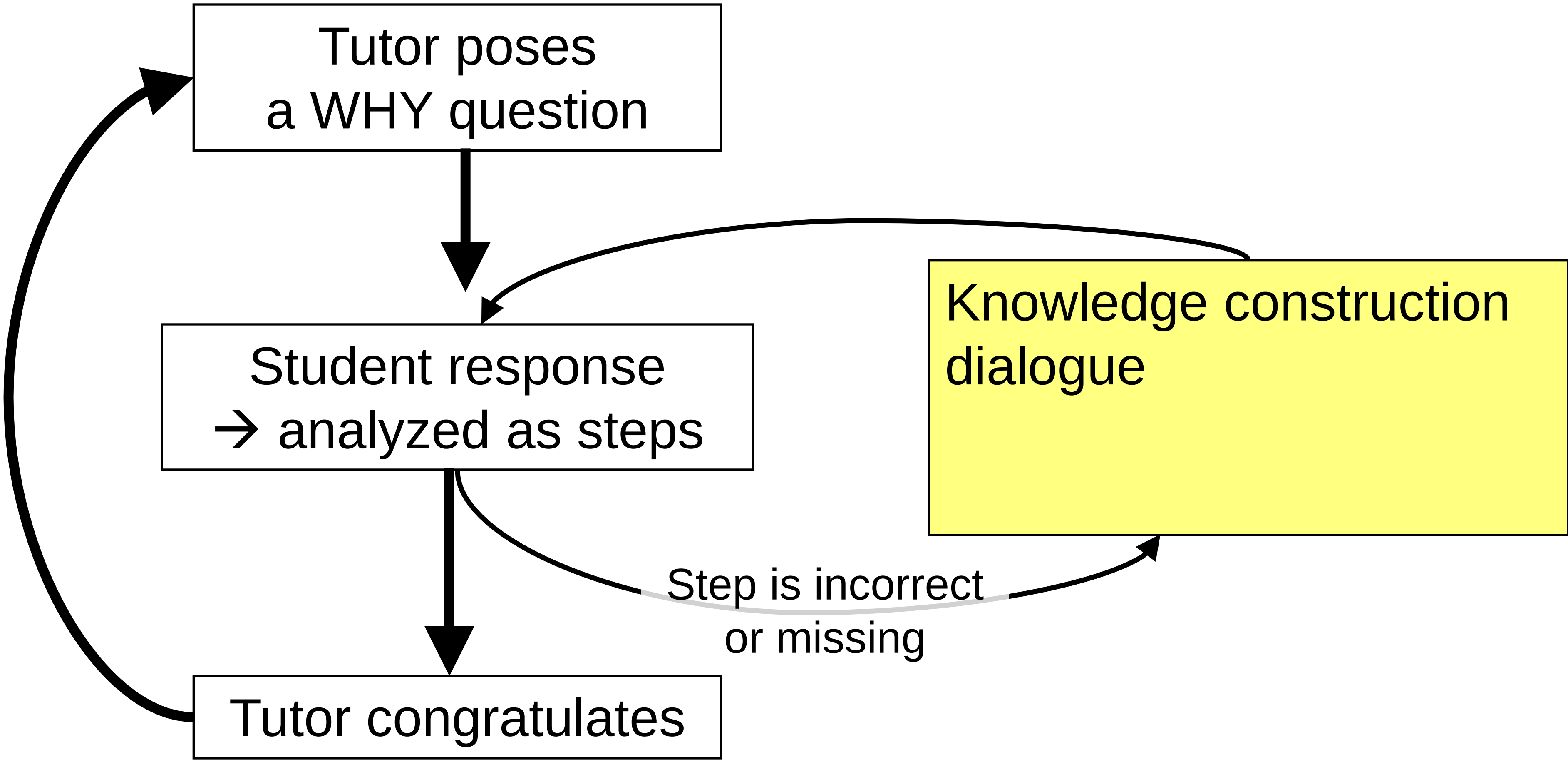
Only difference between tutoring conditions was contents of yellow box



Human tutoring

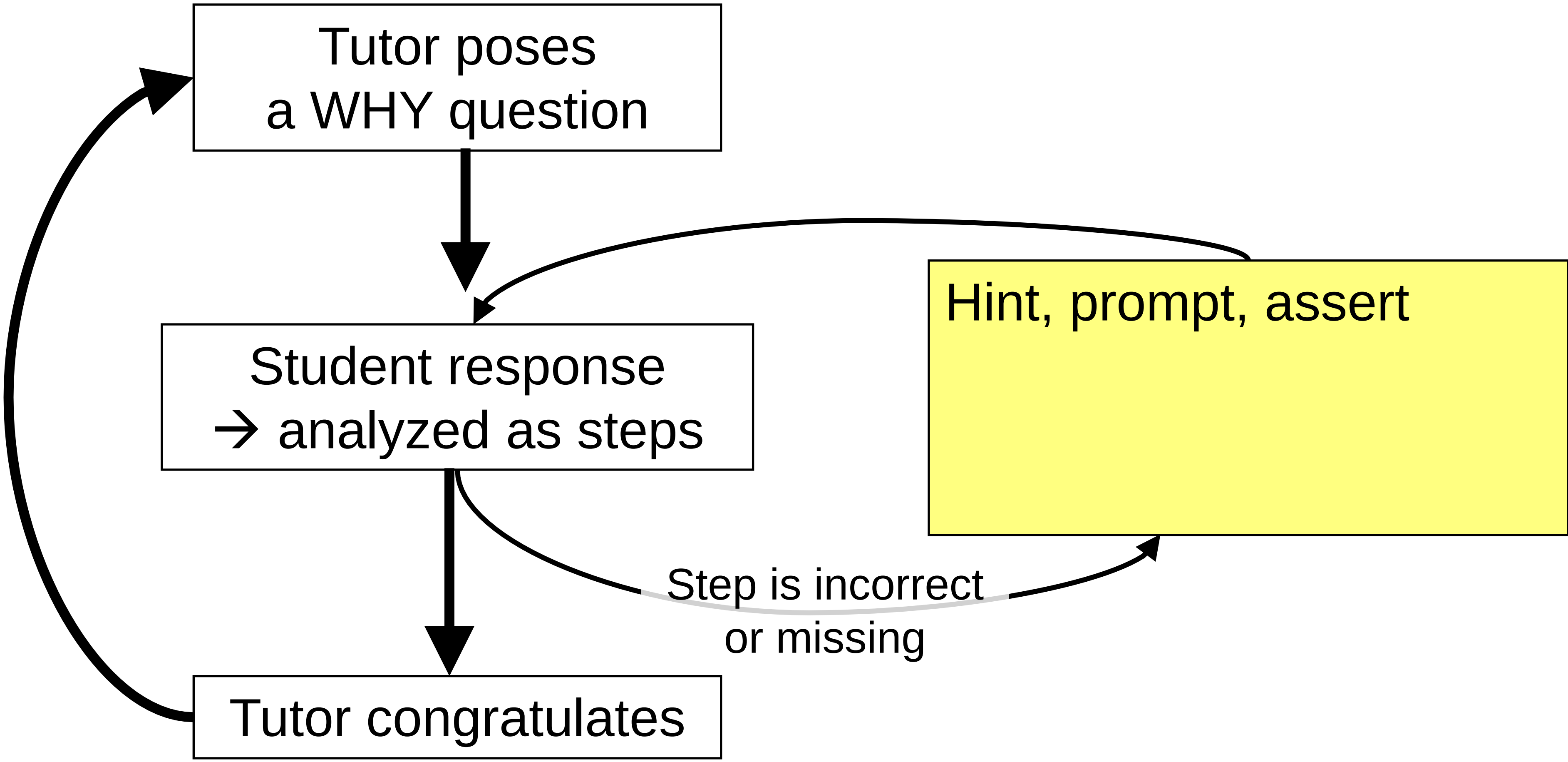


Why2-Atlas

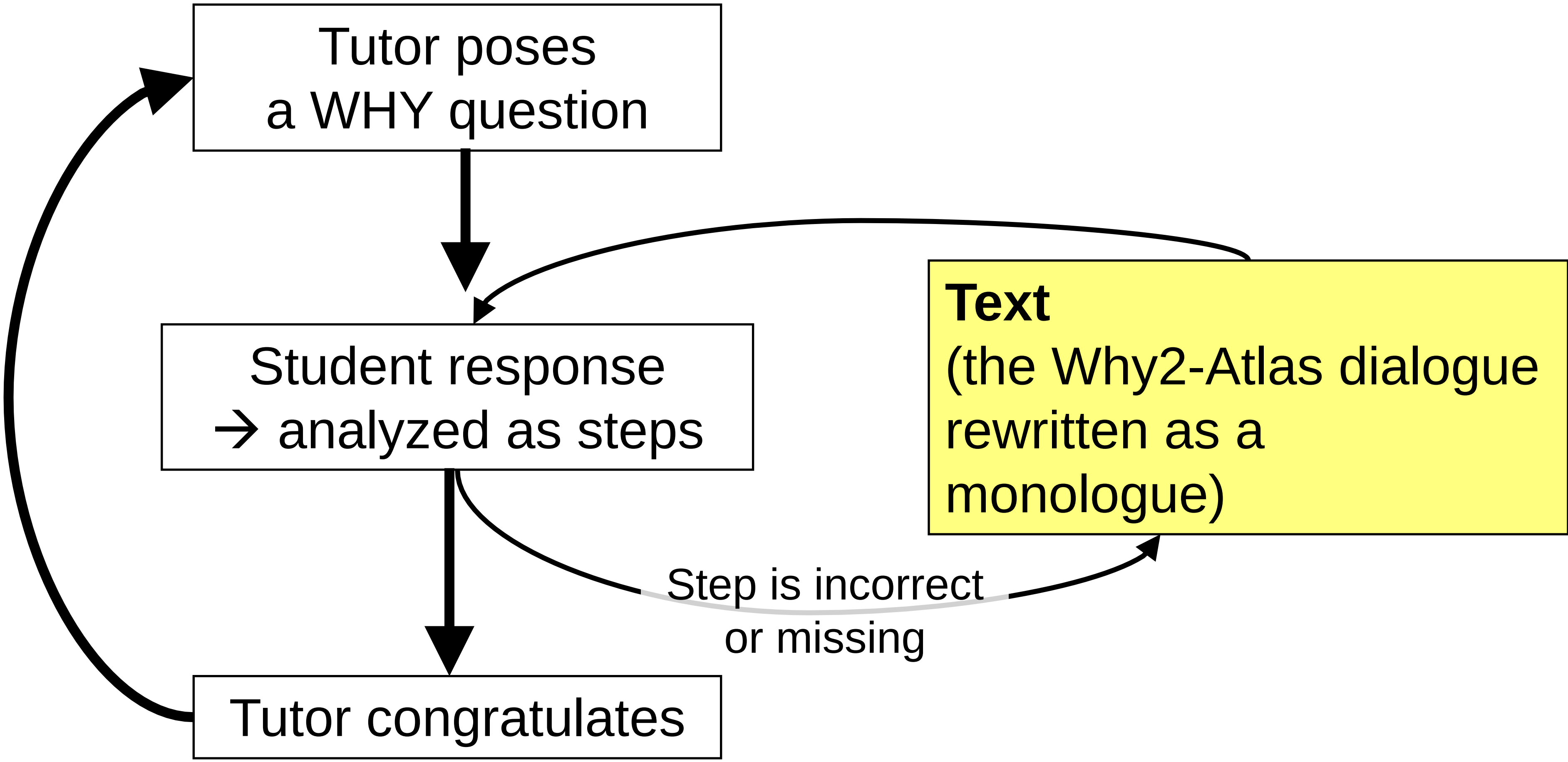


Tutor poses a WHY question
Student response → analyzed as steps
Knowledge construction dialogue
Step is incorrect or missing
Tutor congratulates

Why2-AutoTutor

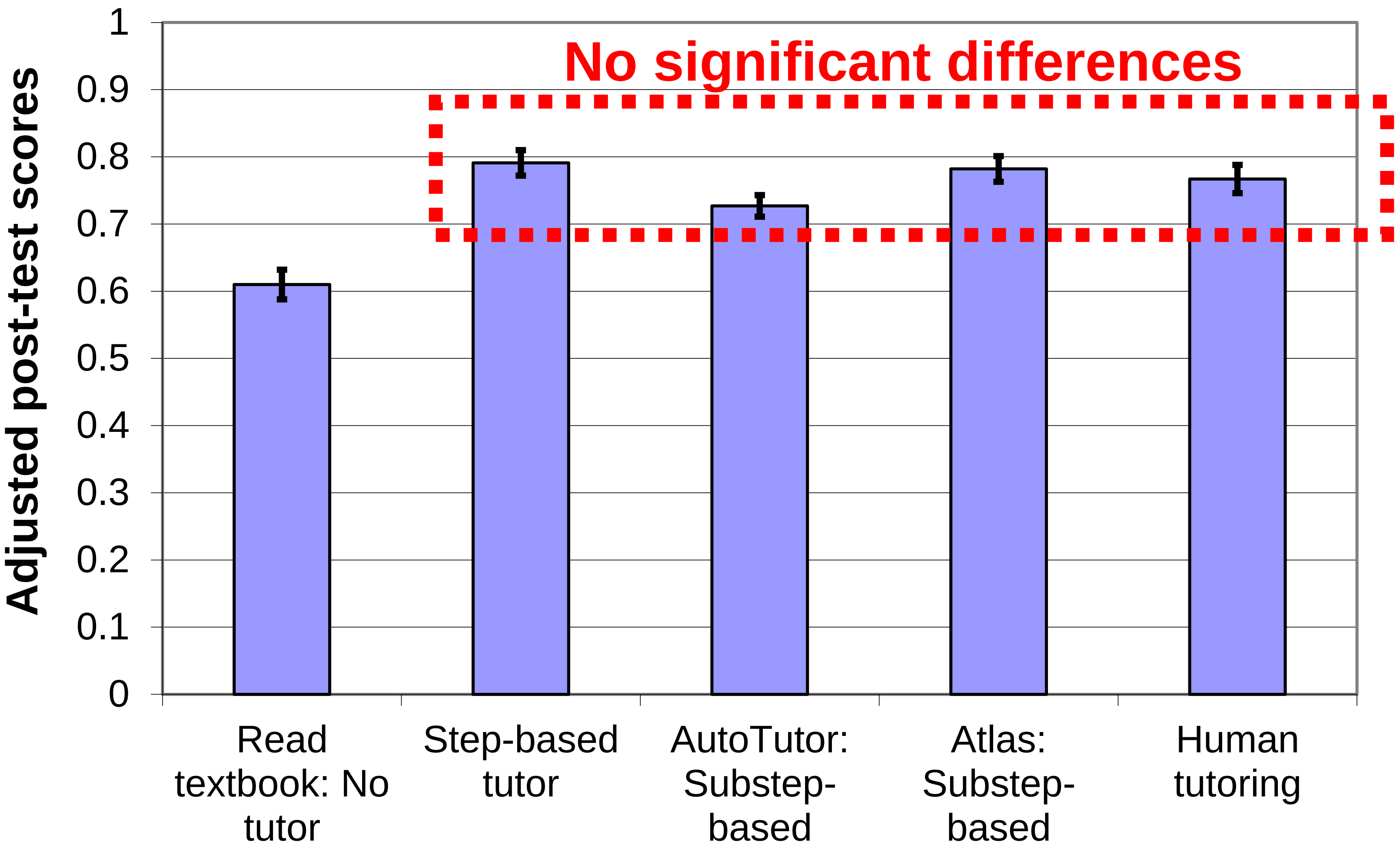


A step-based tutor: A text explanation with same content



Experiments 1 & 2

(VanLehn, Graesser et al., 2007)



Results from all 7 experiments
(VanLehn, Graesser et al., 2007)

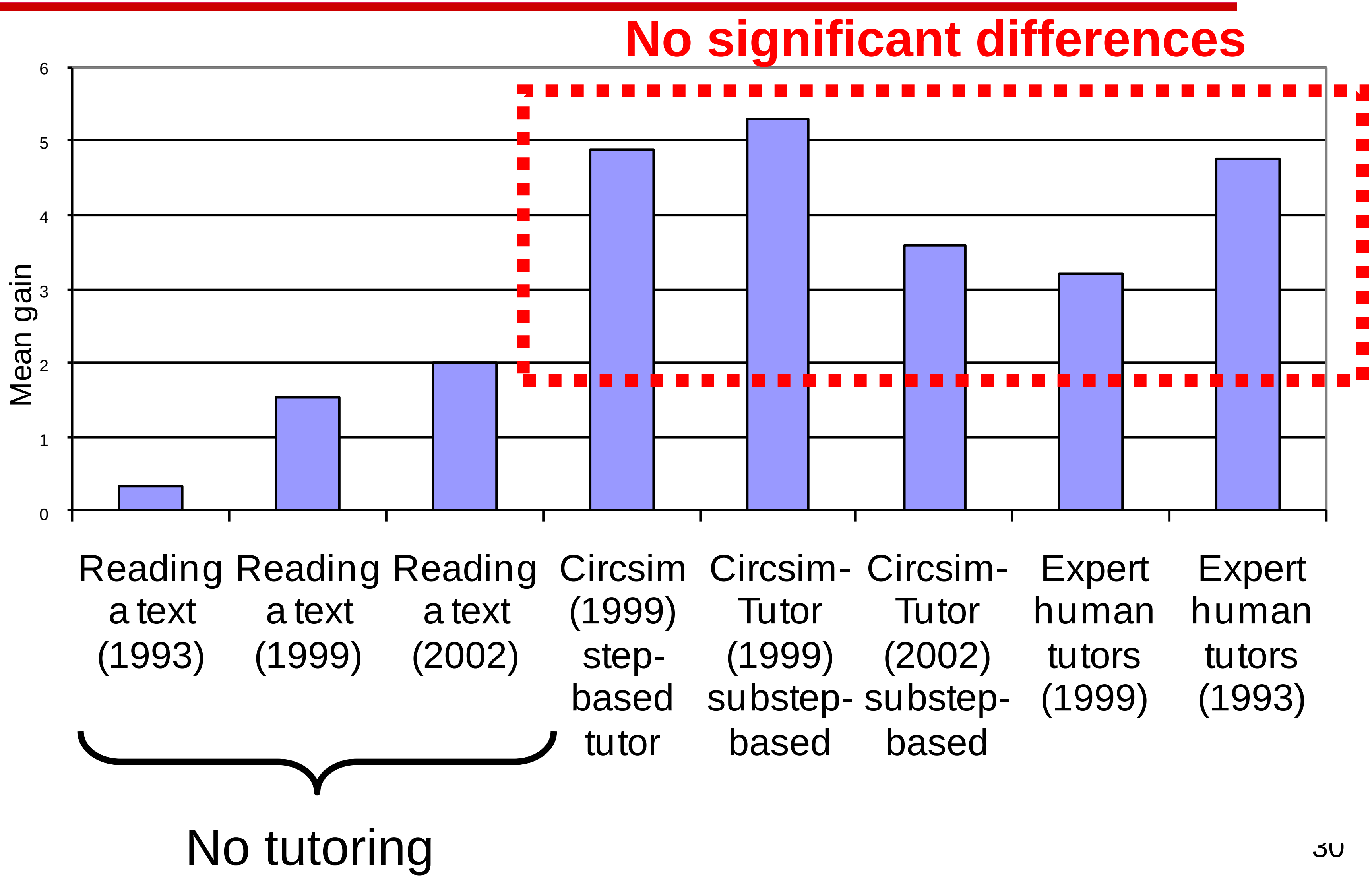
- Human tutoring
  = Substep-based tutoring systems
  = Step-based tutoring system
  - Exception: When pre-physics students worked with instruction authored for post-physics students, then Human tutoring > Step-based tutoring
- Atlas = AutoTutor
- Tutors > Textbook (no tutoring)

Evidence against the interaction granularity hypothesis: Outline

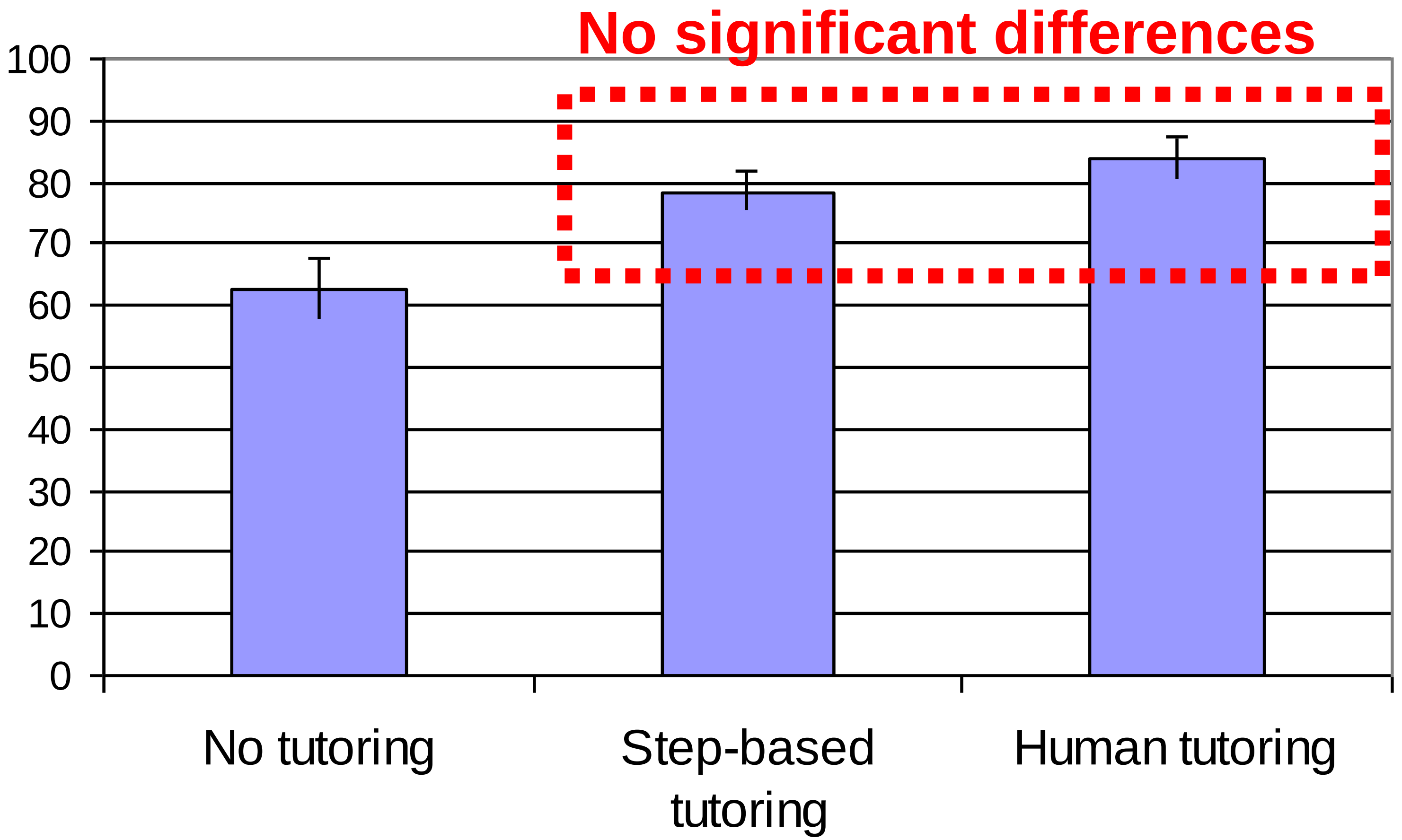
- √ Andes-Atlas
- √ Why2
- ◆ Other studies
- ◆ Meta-analysis

Next

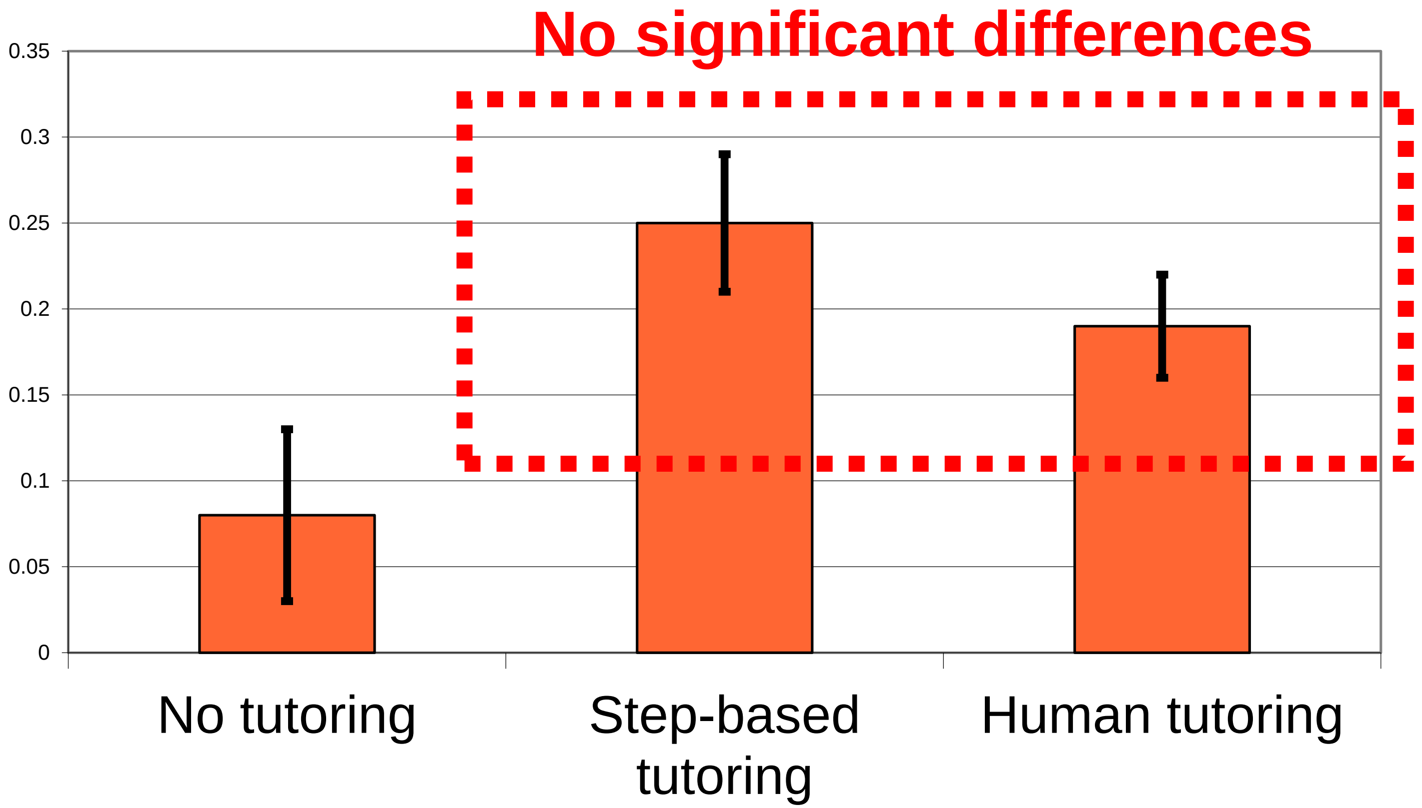
Evens & Michael (2006) also show human tutoring = substep-based tutoring = step-based tutoring



Reif & Scott (1999) also show human tutors = step-based tutoring



Katz, Connelly & Allbritton (2003) post-practice reflection: human tutoring = step-based tutoring



Evidence against the interaction granularity hypothesis: Outline

- √ Andes-Atlas
- √ Why2
- √ Other studies
- ◆ Meta-analysis

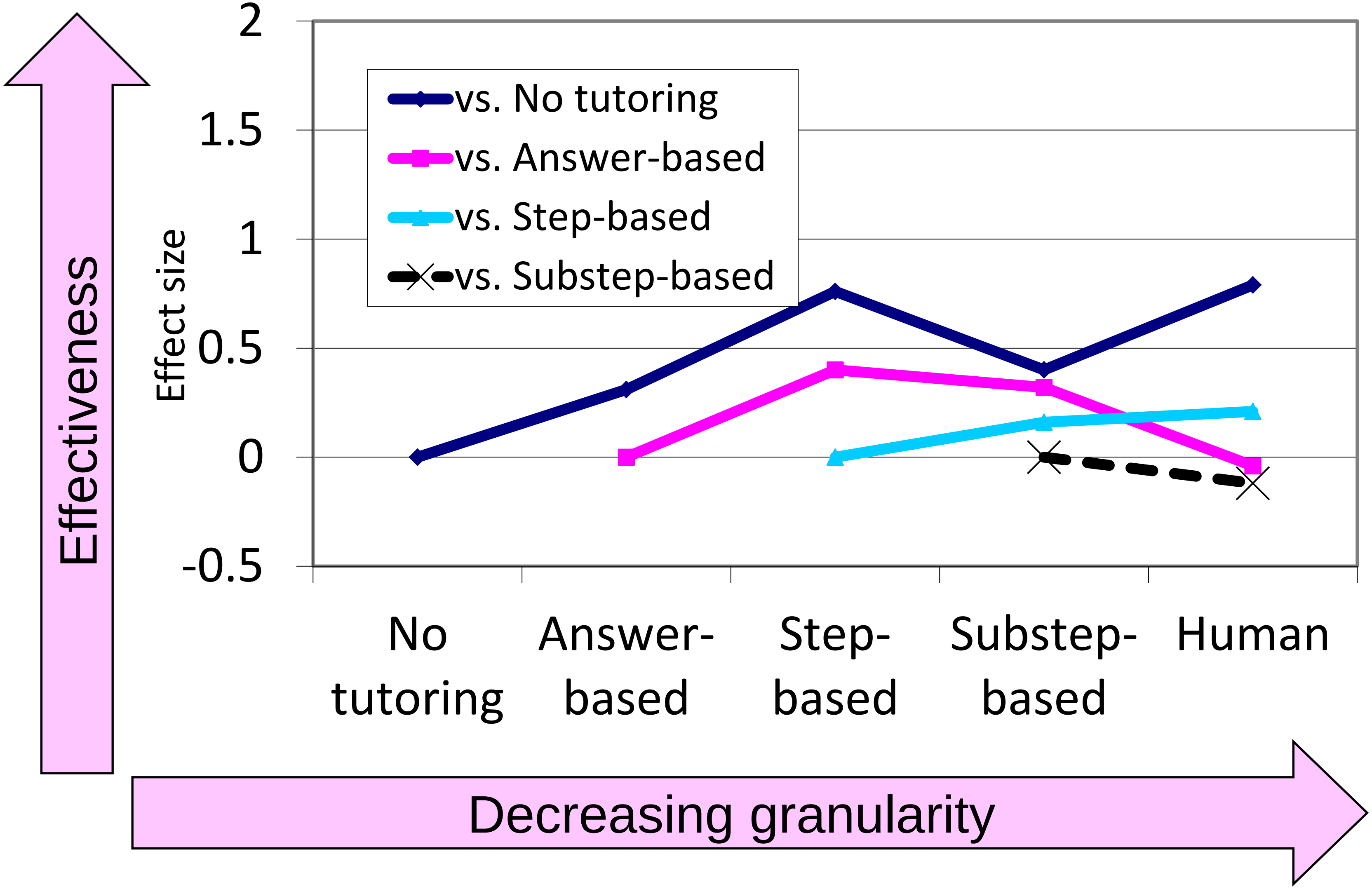
Next

Meta-analytic results for all possible pairwise comparisons (VanLehn, 2011)

Comparison	Num. of effects	Mean effect	% reliable
Answer-based vs. No tutoring	165	0.31	40%
Step-based vs. No tutoring	28	0.76	68%
Substep-based vs. No tutoring	26	0.40	54%
Human vs. No tutoring	10	0.79	80%
Step-based vs. Answer-based	2	0.40	50%
Substep-based vs. Answer-based	6	0.32	33%
Human vs. Answer-based	1	-0.04	0%
Substep-based vs.Step-based	11	0.16	0%
Human vs. Step-based	10	0.21	30%
Human vs. Substep-based	5	-0.12	0%

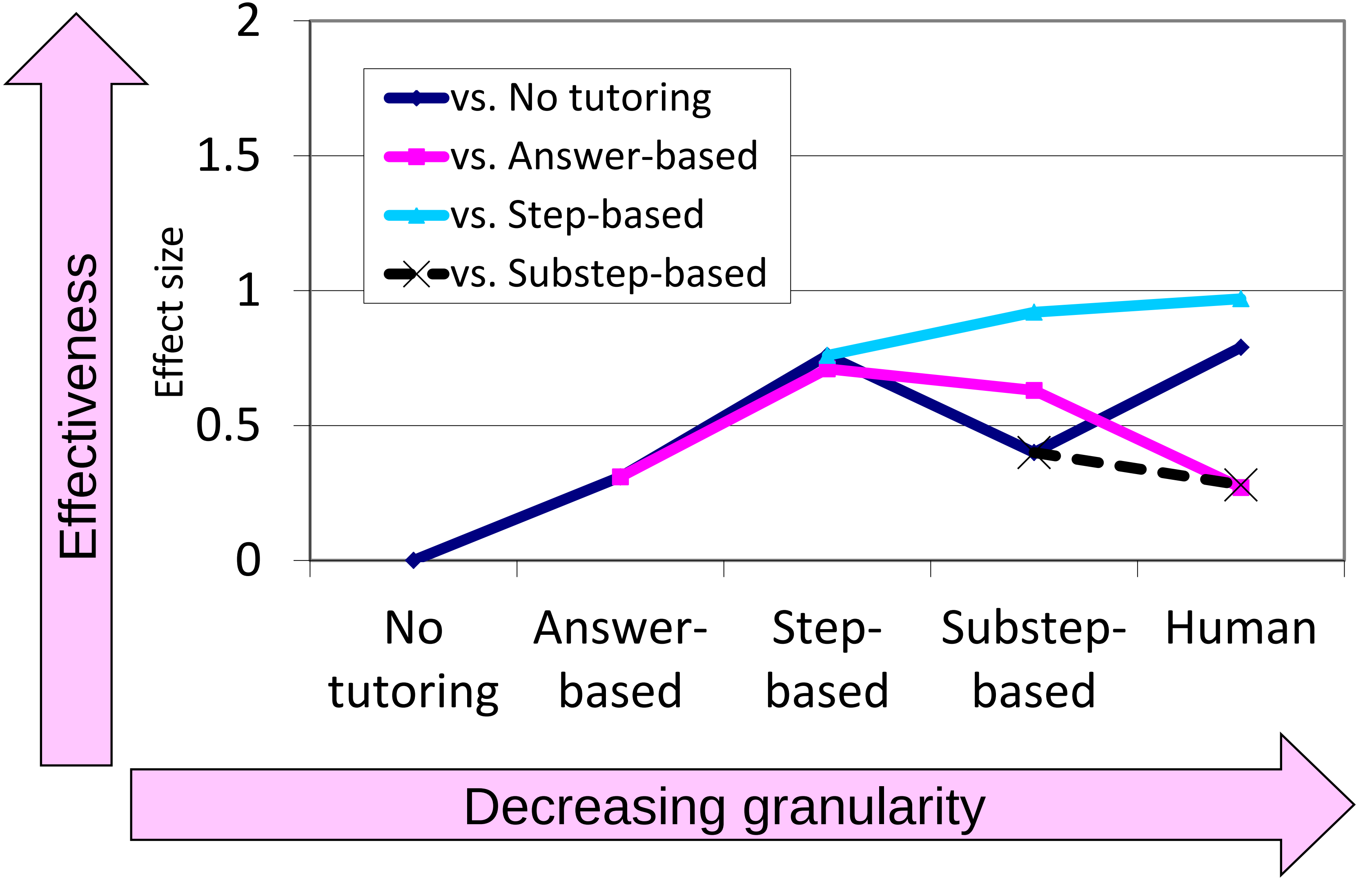
Graphing all 10 comparisons:

graph is hard to understand...

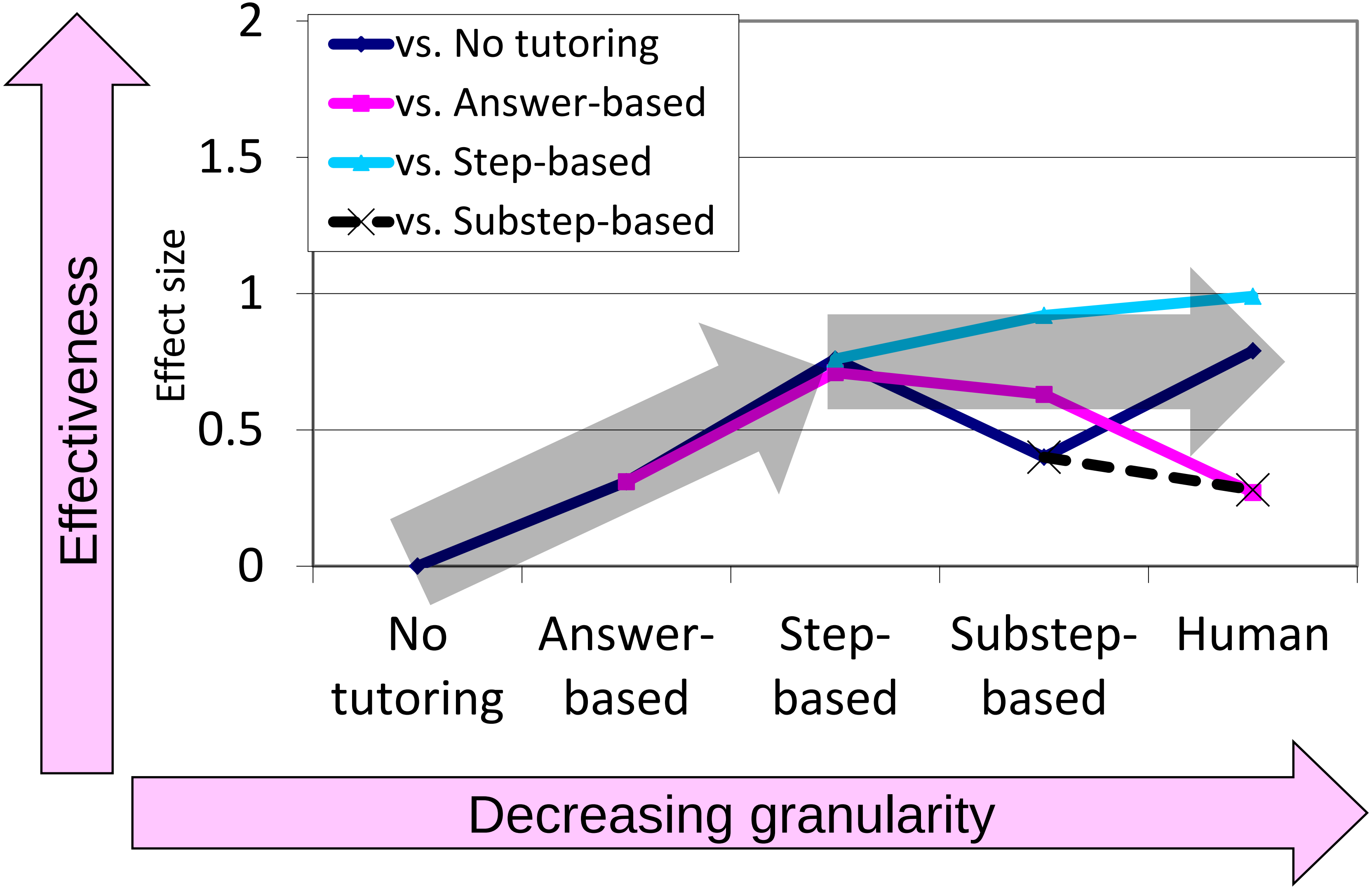


Graphing all 10 comparisons:

Lines raised to make it easier to integrate evidence



The Interaction Plateau Hypothesis:
human = substep = step > answer > none



Outline

- The interaction granularity hypothesis
  - The smaller the grain size of interaction, the more effective the tutoring
  - Grain size: Human < ITS < CAI < no tutoring
  - Effectiveness? Human > ITS > CAI > no tutoring
- Evidence *against* the hypothesis
  - Effectiveness! Human = ITS > CAI > no tutoring
  - The interaction plateau hypothesis
- How to achieve ITS > Human effectiveness

Next

3 recent attempts: Outline

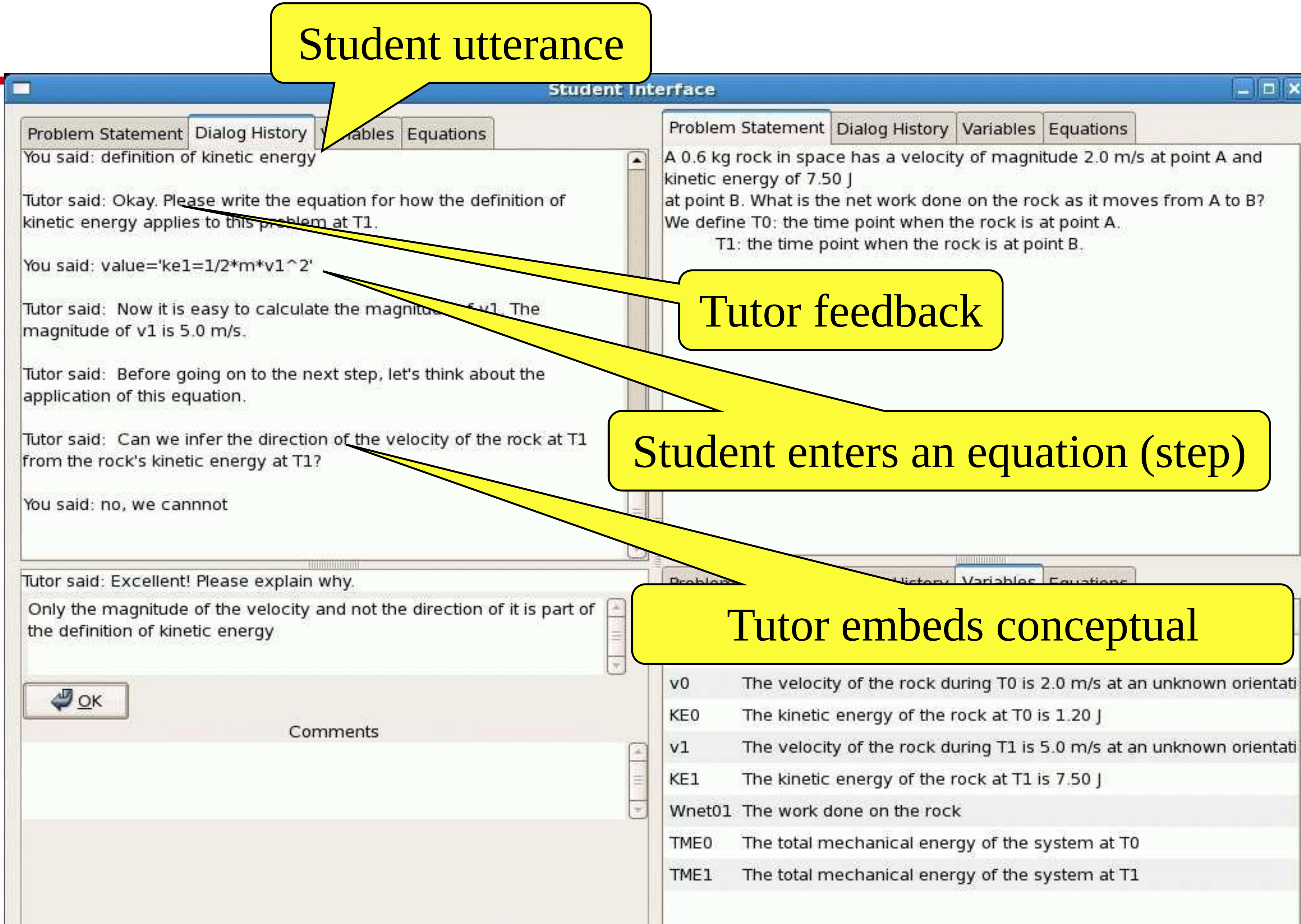
- Embedding conceptual in procedural
- Machine learning of pedagogical tactics
- Meta-strategic scaffolding

Next

Dialogue-based tutoring allows authors to embedded conceptual

- Cordillera is a step-based tutoring system with a natural language dialogue user interface
- Between some steps, it asks conceptual questions that aren’t normally part of the problem solving
  - T: Before going on to the next step, let’s think about the application of this equation. Can we infer the direction of the rock’s velocity at T1 from its kinetic energy?

A dialog-based tutor for physics (Cordillera)



Results

- Cordillera compared to Andes with reflection after problem solving
- For quantitative problem solving, no difference
- For conceptual problem solving, Cordillera > Andes
  - $d=0.50$, $p<.041$
- Interpretation
  - Students probably paid more attention when conceptual instruction was embedded than when it was done afterwards

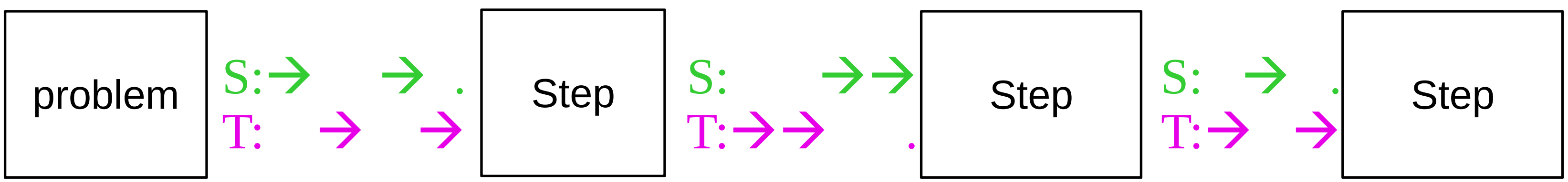
3 recent attempts: Outline

- √ Embedding conceptual in procedural
  - Cordillera produces better conceptual learning (d=0.49) than Andes, and Andes ≈ human
- ◆ Machine learning of pedagogical tactics
- ◆ Meta-strategic scaffolding

Next

A self-improving tutoring system

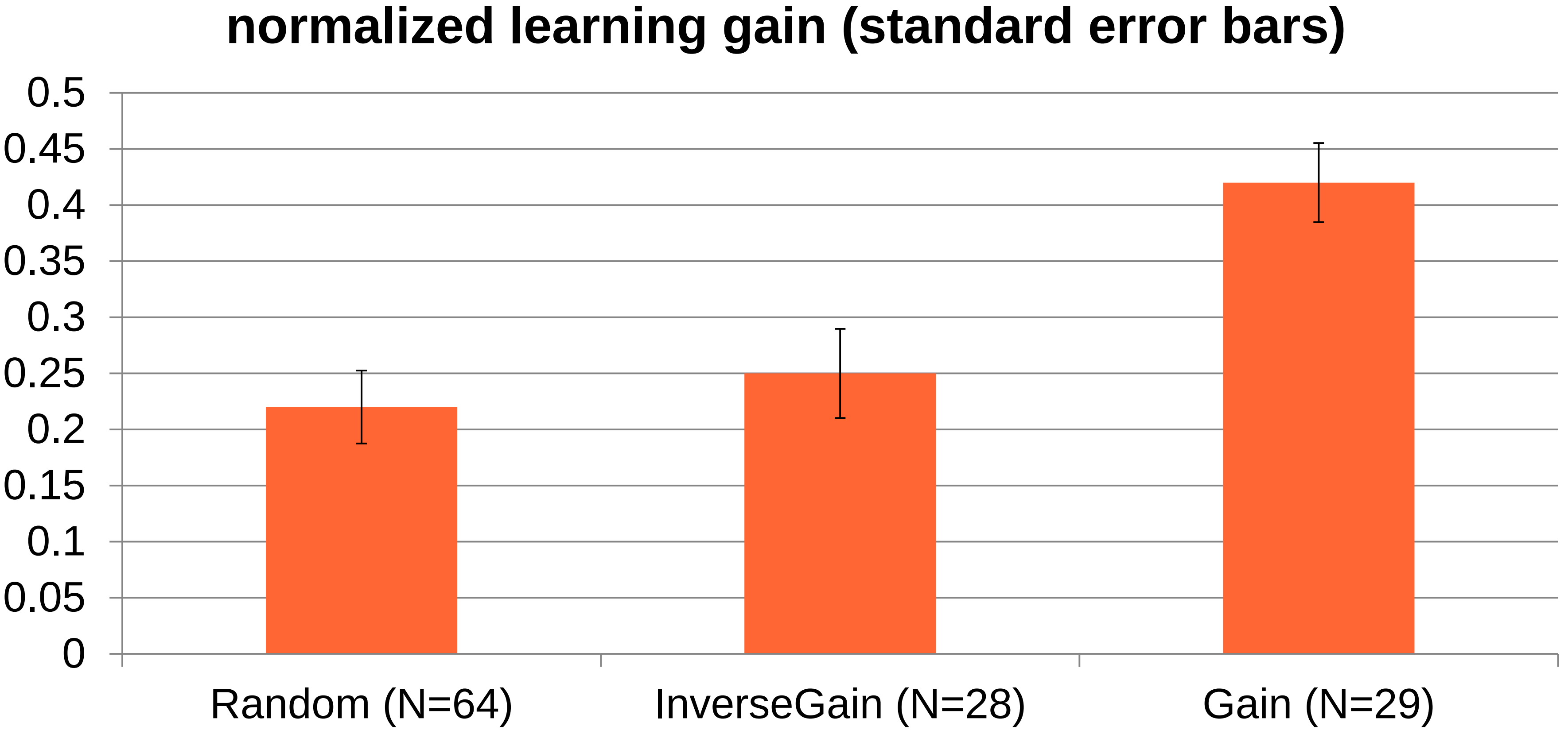
- Dialogue-based physics tutor (Cordillera)
- Chooses between *elicit* and *tell*



- Procedure
  - Collect learning gains using random choice
  - Reinforcement learning, where reward is:
    - » Gain: learning gain
    - » InverseGain: –learning gain
  - Install 2 induced policies in Cordillera
  - Measure learning gains again

Results

Induced policy (Gain) > Random policy by d=0.84, p < .005



3 recent attempts: Outline

- ✓ Embedding conceptual in procedural
  - Cordillera produces better conceptual learning (d=0.49) than Andes, and Andes ≈ human
- ✓ Machine learning of pedagogical tactics
  - Machine learned tactics produced better learning than Cordillera (d=0.84) with random policy
- ◆ Means-ends analysis (MEA) as temporary scaffolding

Next

Means-ends analysis (MEA) is a general problem solving strategy

1. Remove one goal from the set of current goals
2. Select an operator that will achieve or at least partially achieve the target goal.
3. Apply the operator.
4. If this produces new goals, add them to the set of goals.
5. If the set of goals is not empty, go to step 1.

Knowledge base is a set of operators

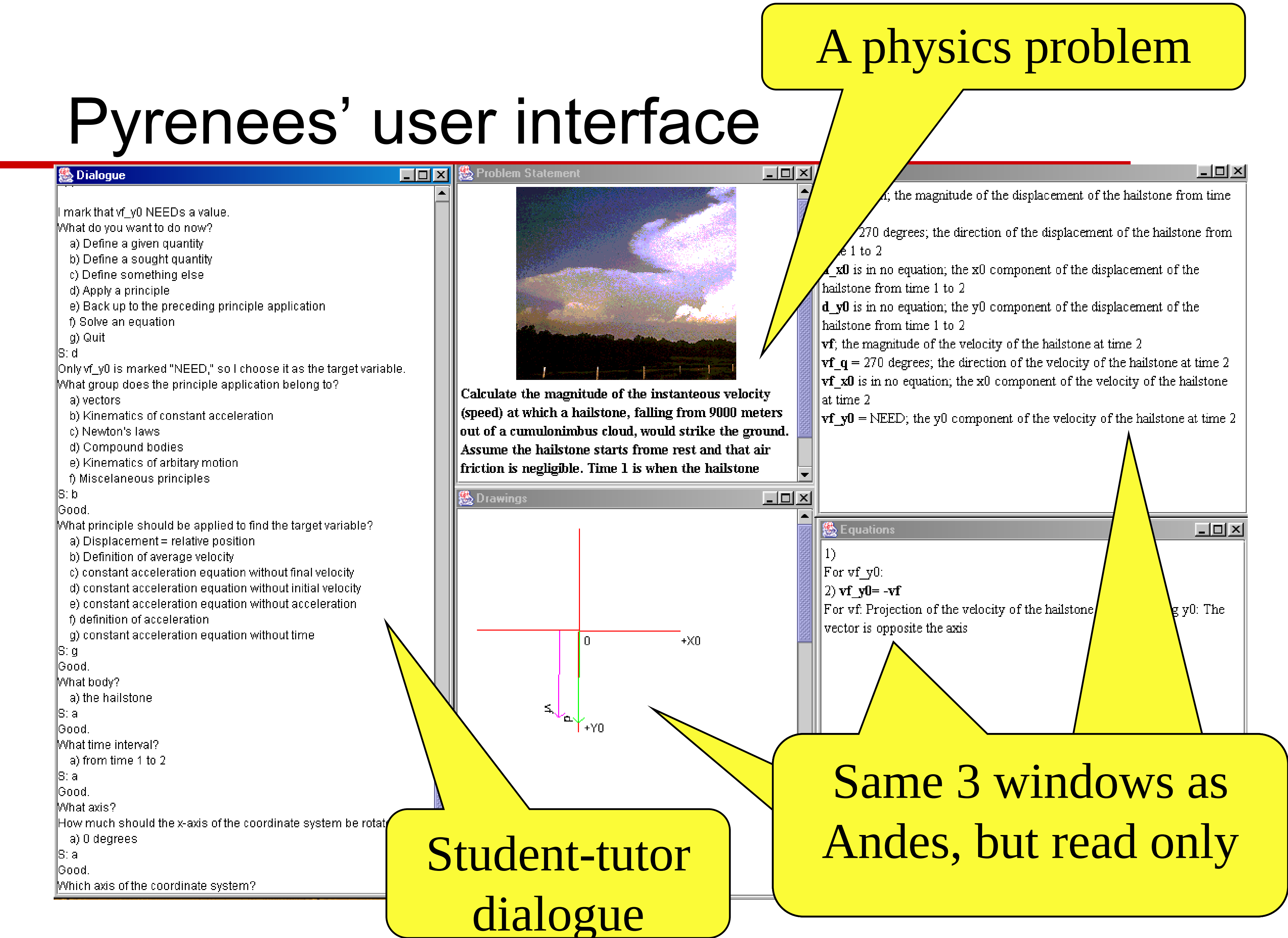
MEA is ...

- Old
  - Aristotle's *Nicomachean Ethics*
  - General Problem Solver (Newell & Simon, 1972)
  - Prolog (Colmerauer, ~1972)
- General
- Tedious
- Used by neither experts nor novices
  e.g., when solving physics problems
  - Simon & Simon (1978), Larkin (1983), Priest (1992)
- Not taught
  - Our physics instructors refused

Teaching MEA as temporary scaffolding might help learning.

- Knowledge is taught as a set of operators
  - In physics, operators = principles
  - Famous principles include Newton's second law...
  - Tedious principles include V_X= V cos(V_θ) ...
- Students are initially required to use MEA
- But then use of MEA is optional
  - Probably will use it only when they get stuck
- Would MEA transfer to a new task domain?

Pyrenees' user interface



Pyrenees requires that students follow a specific strategy

- Andes does not teach a problem solving strategy
  - students tend to copy examples
- Pyrenees teaches a general problem solving strategy
  - Remove a variable from set of goals
  - Select **principle** that could contain the variable
  - Apply the principle, generating an equation
  - If the equation has any unknown variables, then add them to goals
  - Repeat until no goals left

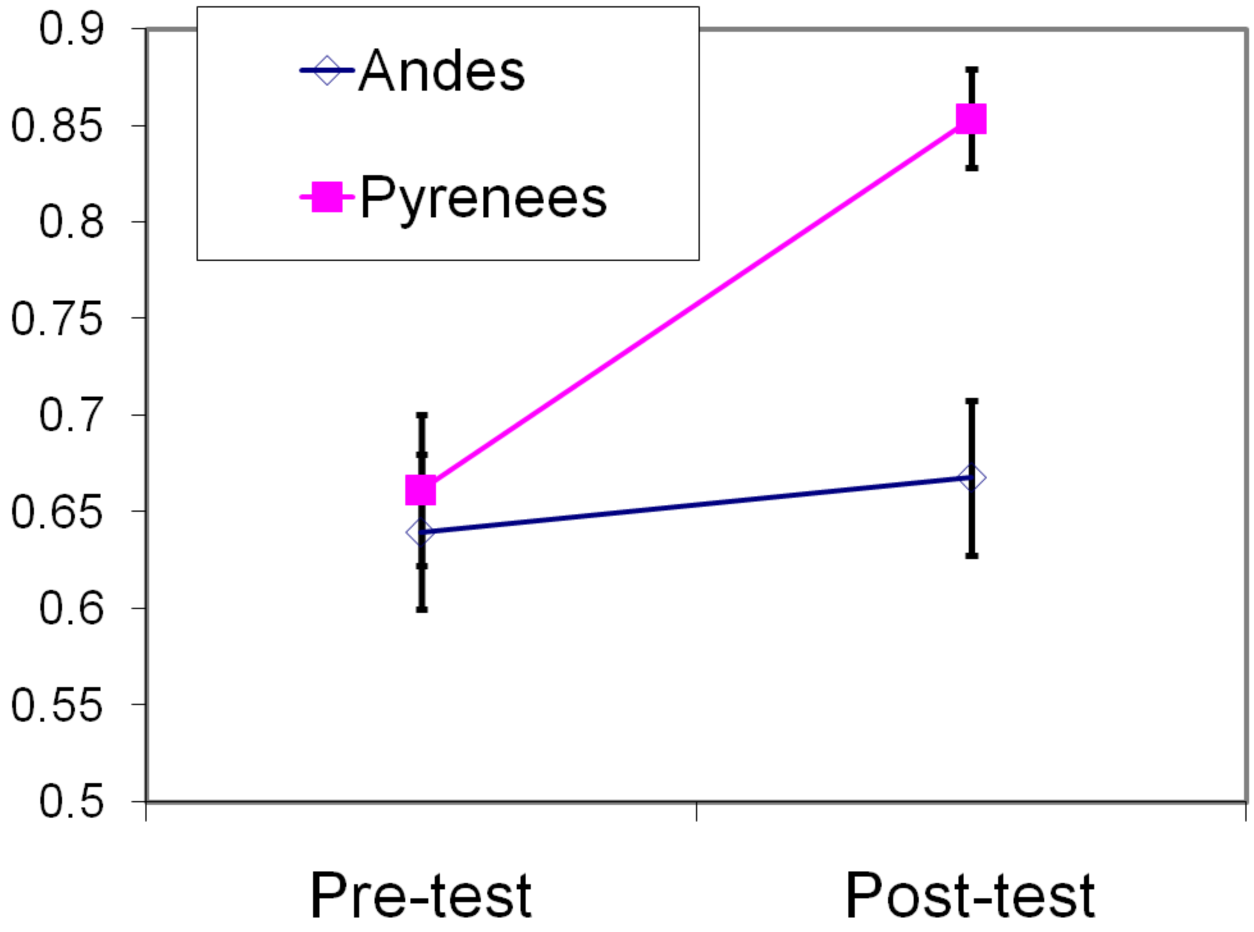
Know thy principles!

Experimental Procedure

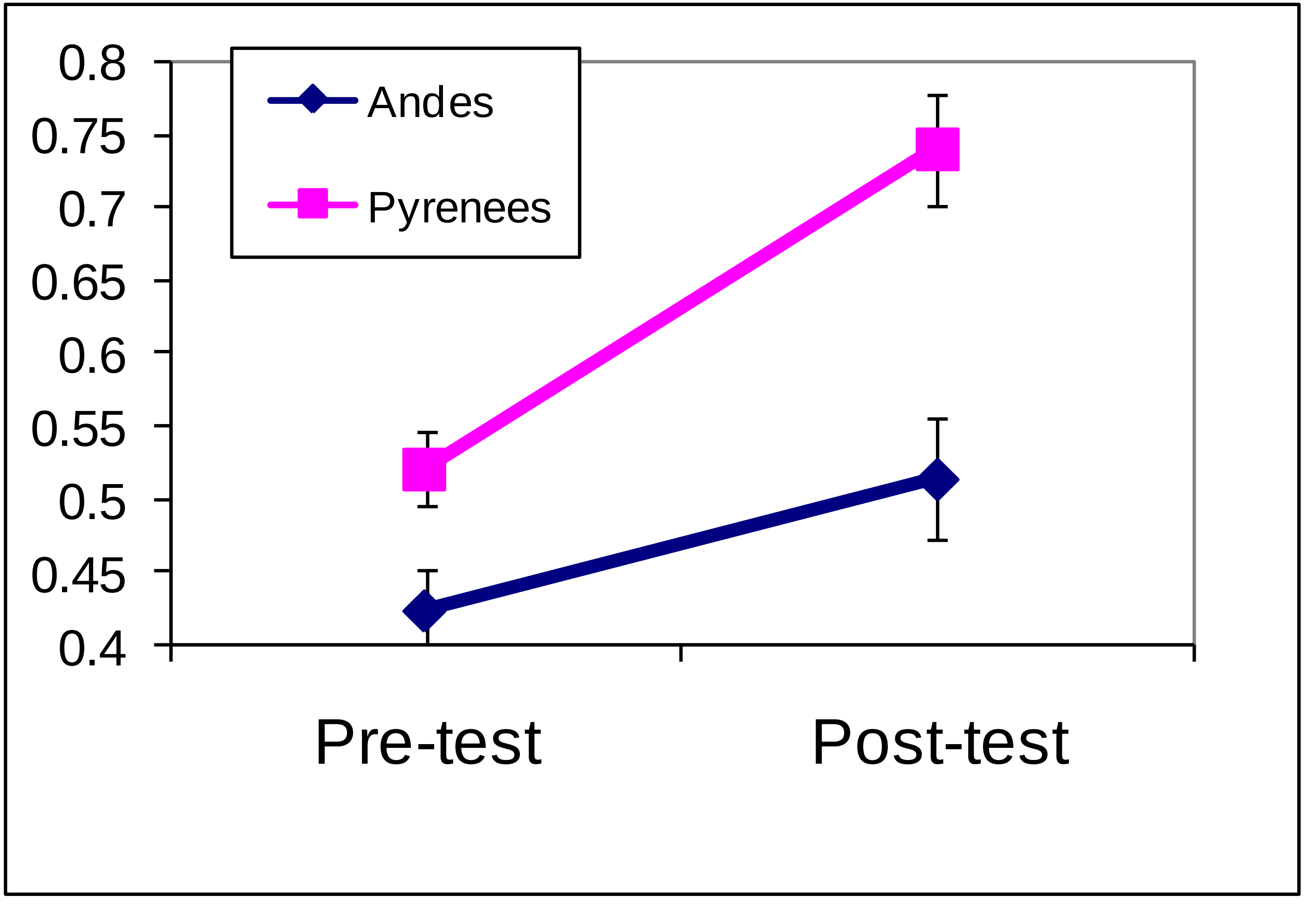
	Pyrenees group	Andes group
Probability Instruction	Pyrenees	Andes
Physics Instruction	Andes	

Instruction =
pre-training → pre-test → training → post-test

Results from initial domain (probability) d = 1.17 (post)



Results from second domain (physics) d = 0.69 (pre) d=1.28 (post)



ITS can be improved, and may now be more effective than humans

- ✓ Embedding conceptual in procedural
  - Cordillera produces better conceptual learning (d=0.49) than Andes, and Andes ≈ human
- ✓ Machine learning of pedagogical tactics
  - Machine learned tactics produced better learning than random-policy Cordillera (d=0.84)
- ✓ Means-ends analysis (MEA) as temporary scaffolding
  - Produced better learning (d=1.17) than Andes
  - Produced better learning in a second task domain (d=1.28) where it was not explicitly taught

Why will ITS eventually become more effective than human tutors?

- Innovative instruction (see 3 preceding examples)
- Quality assurance
  - Human tutors make many mistakes
  - Step-based tutors do too, but they can be improved
    - » via a manual Quality Assurance process
    - » via reinforcement learning & other machine learning
- ITS excel at
  - Large library of tasks → adaptive task selection
  - High accuracy stealth assessment
  - Monitoring the student's affective state
- But: Rapport? Off topic discussions?

Questions? (outline below)

- The interaction granularity hypothesis
  - Grain size: Human < ITS < CAI < no tutoring
  - Effectiveness? Human > ITS > CAI > no tutoring
- Evidence *against* the hypothesis
  - Effectiveness! Human = ITS > CAI > no tutoring
  - The interaction plateau hypothesis
- How to achieve ITS > Human effectiveness
  - Innovative instruction
  - Quality assurance
  - Adaptive task selection, stealth assessment, affect monitoring...

Bibliography
(all papers available from public.asu.edu/~kvanlehn)

- The meta-analysis

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

- Why2 experiments

  - VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rose, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science, 31(1), 3-62.*

- Andes, the physics tutor

  - 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.*

- Andes-Cordillera study
  - in prep

Bibliography continued

- Andes-Atlas studies
  - Siler, S., Rose, C. P., Frost, T., VanLehn, K., & Koehler, P. (2002, June). *Evaluating knowledge construction dialogues (KCDs) versus minilesson within Andes2 and alone. Paper presented at the Workshop on dialogue-based tutoring at ITS 2002, Biaritz, France.*
- Machine learning of Cordillera policies
  - Chi, M., VanLehn, K., Litman, D., & Jordan, P. (2011). Empirically evaluating the application of reinforcement learning to the induction of effective and adaptive pedagogical strategies. *User Modeling and User-Adapted Interaction, 21(1-2), 99-135.*
- *Teaching MEA*
  - Chi, M., & VanLehn, K. (2010). Meta-cognitive strategy instruction in intelligent tutoring systems: How, when and why. *Journal of Educational Technology and Society, 13(1), 25-39.*