

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

Kurt VanLehn

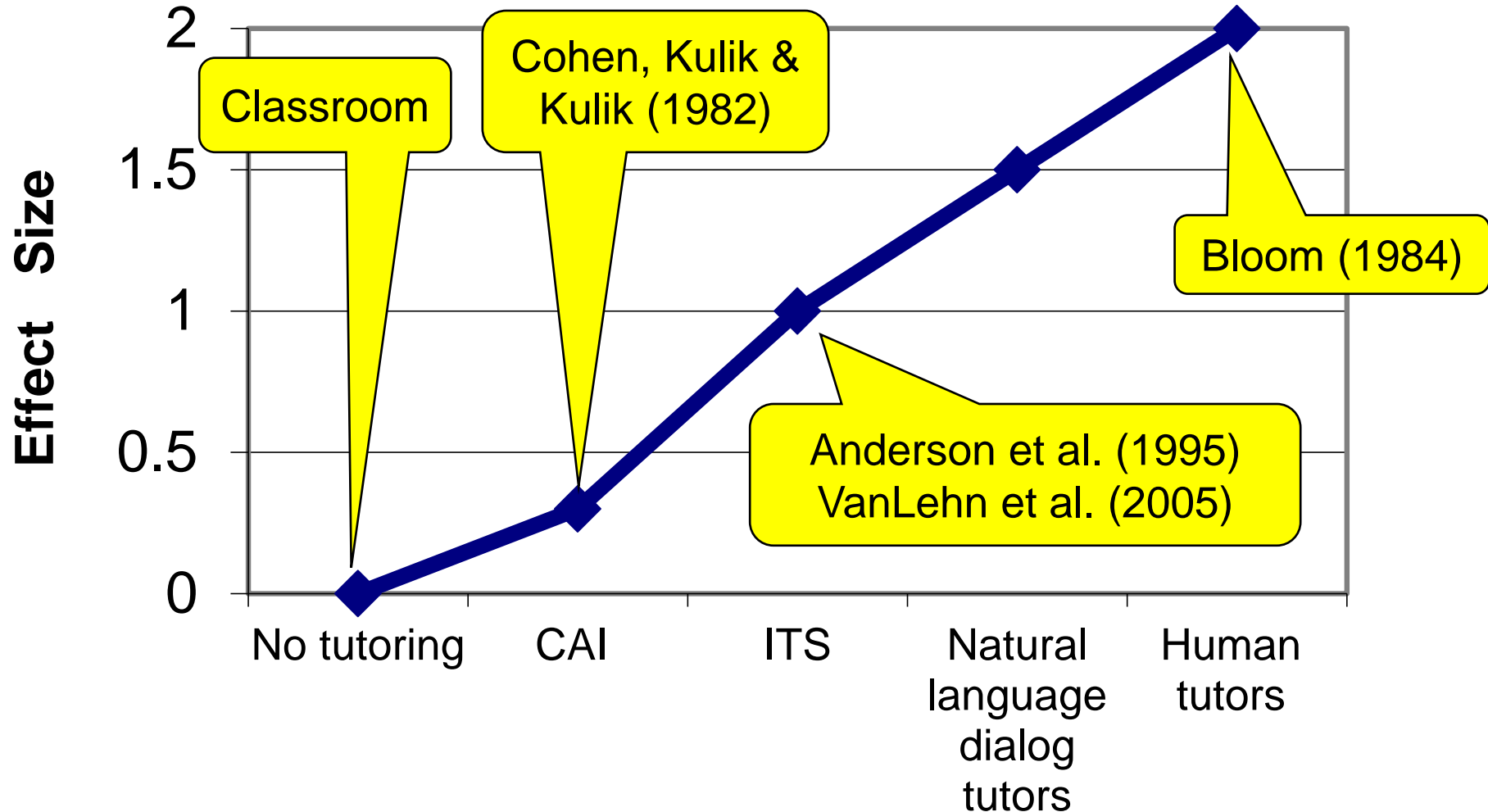
School of Computing, Informatics and Decision Systems Engineering
Arizona State University



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

The screenshot shows the MasteringPHYSICS interface. The top navigation bar includes 'Intro', 'Problem Library', 'Tutorials' (highlighted), 'Gradebook', and 'Reliability'. The main content area is titled 'Conical Pendulum I'. It contains a text description of the problem: 'A bob of mass m is suspended from a fixed point with a massless string of length L (i.e., it is a pendulum). You are to investigate the motion in which the string moves in a cone with half-angle θ .' Below the text is a diagram of a conical pendulum with a bob of mass m , string length L , and angle θ . The interface is divided into 'Part A' and 'Part B'. Part A asks for the tangential speed v in terms of m , L , θ , and g . The answer box shows the expression $v = L \cdot g \cdot \sin(\theta) \cdot \tan(\theta)$. Part B asks for the time for one full revolution. The answer box is empty. Navigation buttons like 'submit', 'hints', 'show answer', and 'review part' are visible below the answer boxes.

Part A

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

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

$v =$

$L \cdot g \cdot \sin(\theta) \cdot \tan(\theta)$

Part B

How long does it take the bob to make one full revolution (one complete trip around the circle)?

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

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

Hints

“Bottom out” hint
i.e., correct answer

An ITS (Andes) with WIMP interface

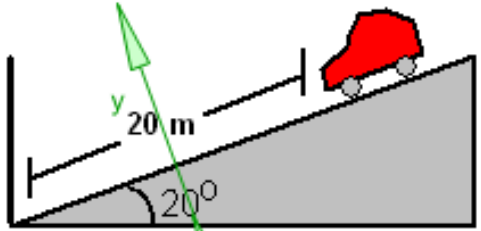
→ Step-based tutoring

dt5a Edit Physics Help

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 is the magnitude of the car's velocity when it hits the garage door?

Answer:



Let N be the normal force due to the driveway

Let F_w be force of gravity

Let a be acceleration of the car

Time T0: car starts rolling.

Time T1: car hits garage door.

g is the acceleration of gravity on earth
 $g = 9.8 \text{ m/s}^2$

Let m be the mass of the car
 $m = 2000 \text{ kg}$

$F_w_x =$

Type in equation (a step)

Draw axes (a step)

Draw vector (a step)

Natural language (NL) dialogue tutoring

→ substep based tutoring

Question

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 essay

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?

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.

If a step in the essay's argument is missing or incorrect, discuss it here.

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 $P(\dots)$?

Hint

S: $P(\text{worm} \& \text{rot})$.

T: Almost. Check your "rot"

Negative feedback

S: $P(\text{worm} \& \sim \text{rot})$

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

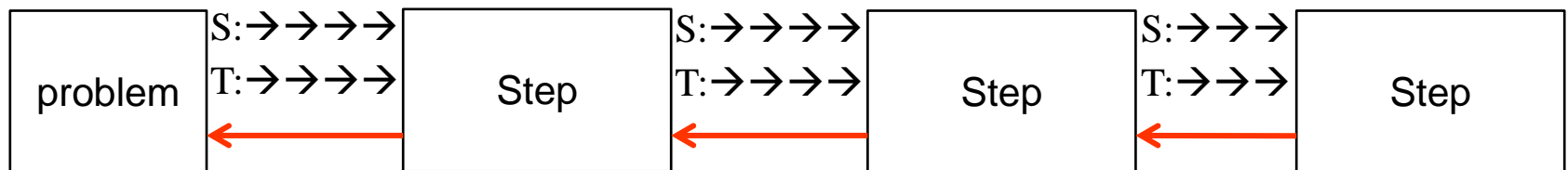
S: $P(A \& B) = P(A) * P(B)$

Granularity of tutoring \approx number of inferences (\rightarrow) between interactions

◆ Answer-based tutoring (CAI)



◆ Step-based tutoring (ITS with WIMP)



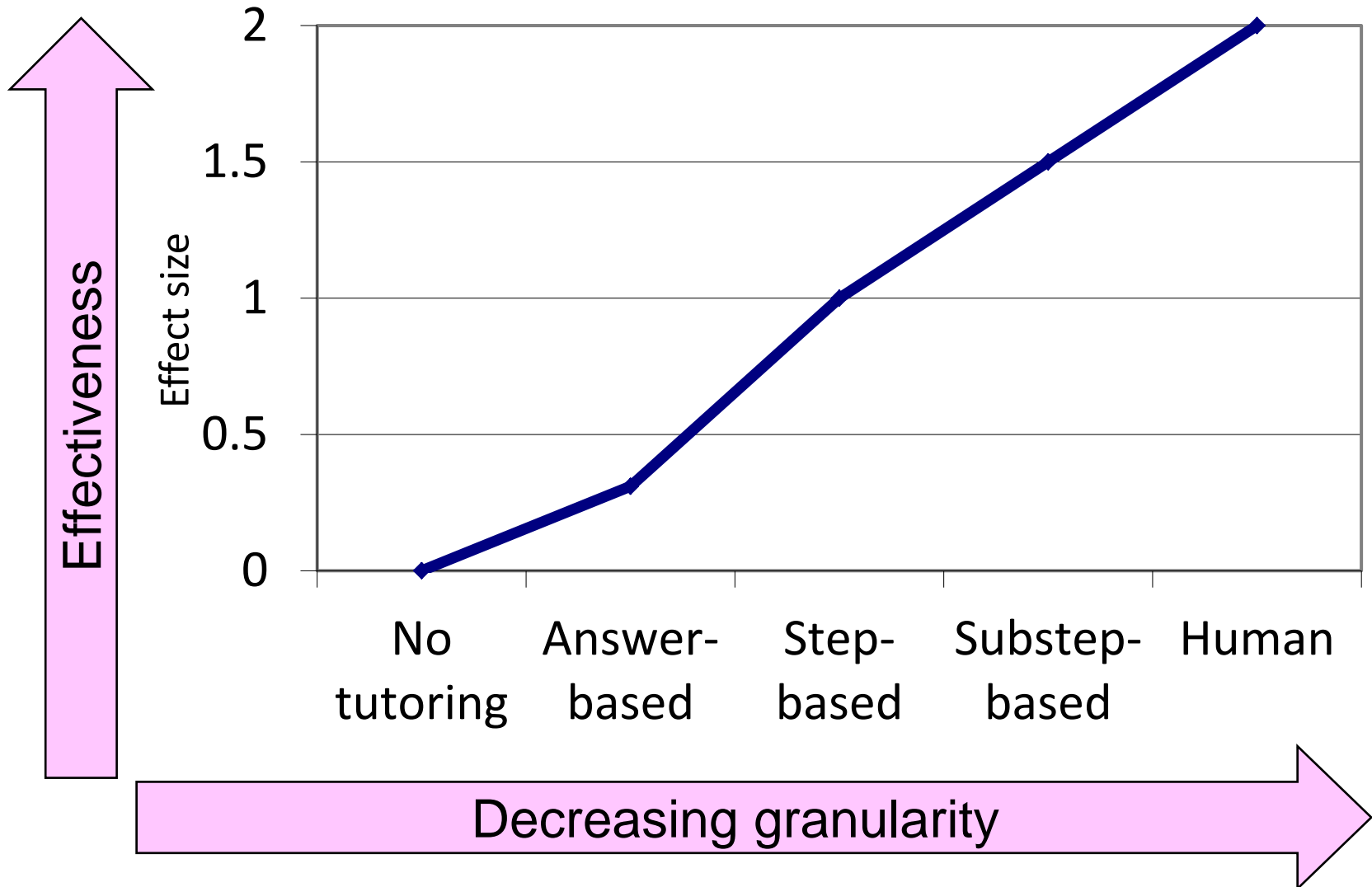
◆ Human tutoring



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
 - Effectiveness! Human = ITS > CAI > no tutoring
 - The interaction plateau hypothesis
- ◆ How to achieve ITS > Human effectiveness



Next

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 \approx Andes ($p > .10$)
- ◆ Study 3 (N=21): Andes-Atlas < Andes ($p < .10$, $d = 0.34$)
- ◆ Study 4 (N=12): Andes-Atlas \approx 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

The screenshot shows a web browser window titled "Netscape: Interactive Conceptual Tutoring". The interface includes a menu bar (File, Edit, View, Go, Window, Help) and a main content area. A problem statement is displayed at the top right: "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?". Below this, there are two main text areas. On the left, under the heading "Dialogue:", is a log of the conversation. On the right, under the heading "Enter your essay here:", is a text input area for the student's response. At the bottom left, there is a text input field labeled "Tell Tutor:" and a button labeled "Send Essay or Message".

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

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's essay

Enter your essay here:

At the very top of the path, where the velocity is equal to 0, the acceleration will be $-9.8 \text{ m/(s}^2\text{s)}$, 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.

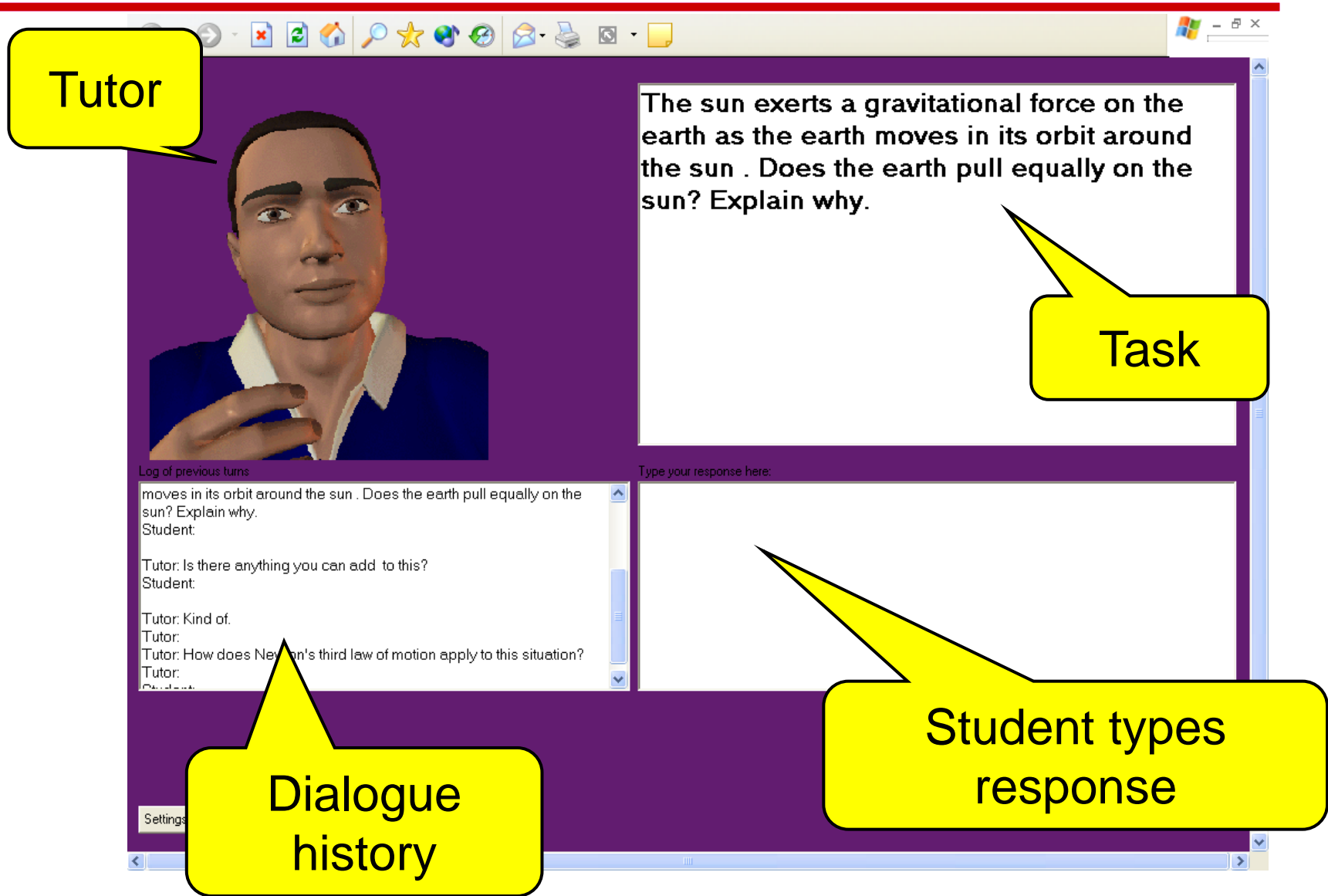
Student's turn in the dialogue

Tell Tutor:

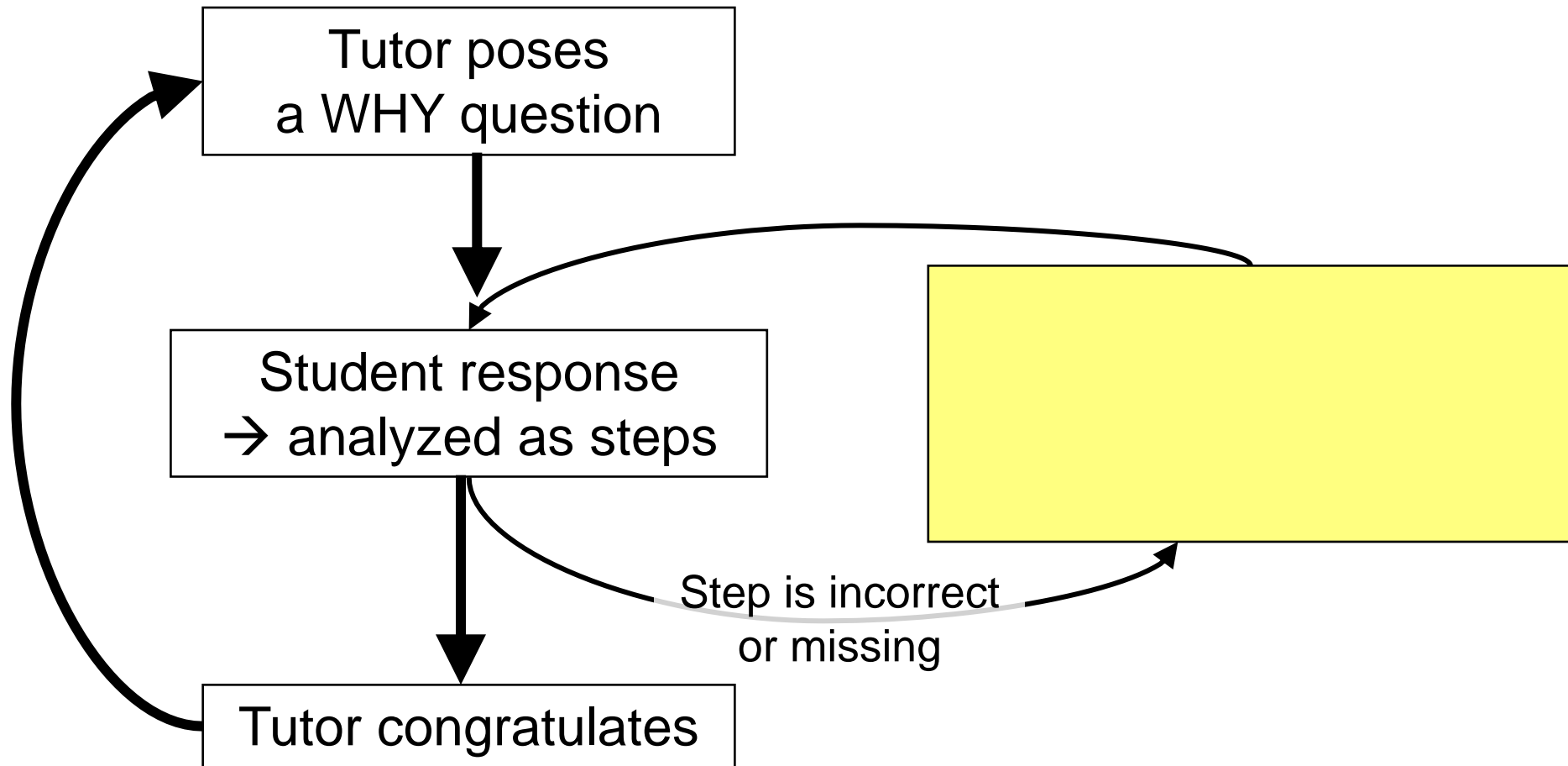
They will be equal. The height the ball reaches will depend on the initial velocity.

Send Essay or Message

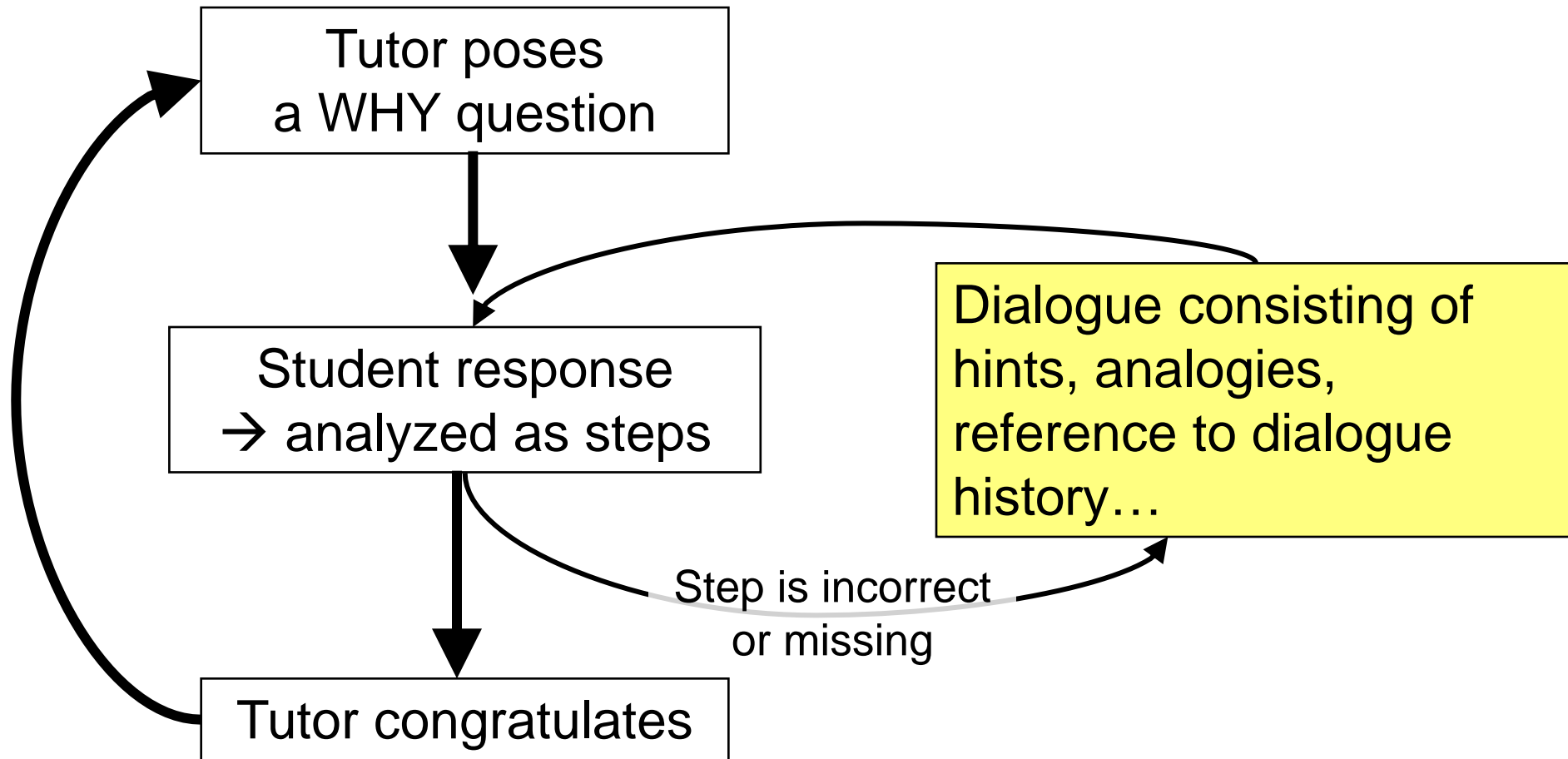
Why2-AutoTutor user interface



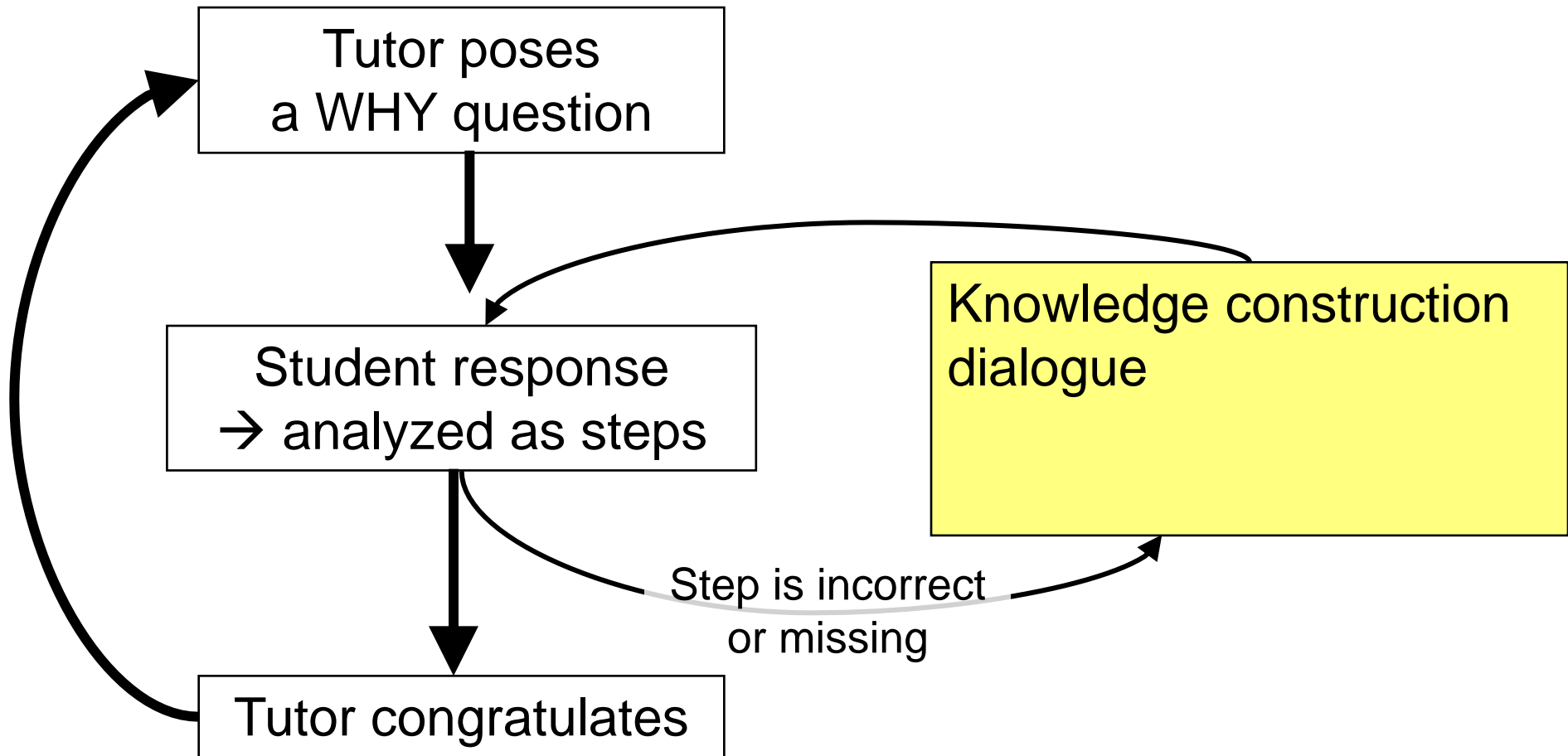
Only difference between tutoring conditions was contents of yellow box



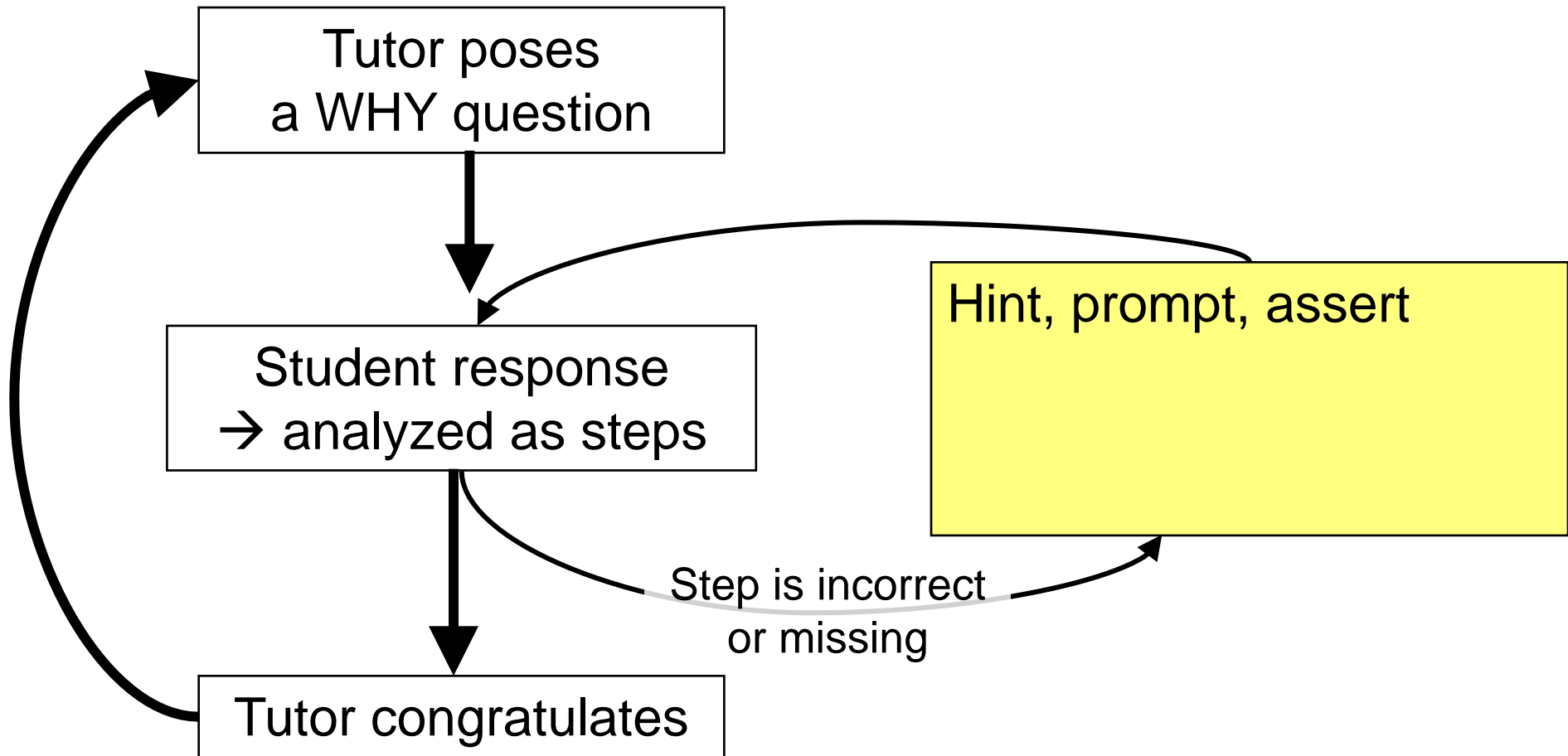
Human tutoring



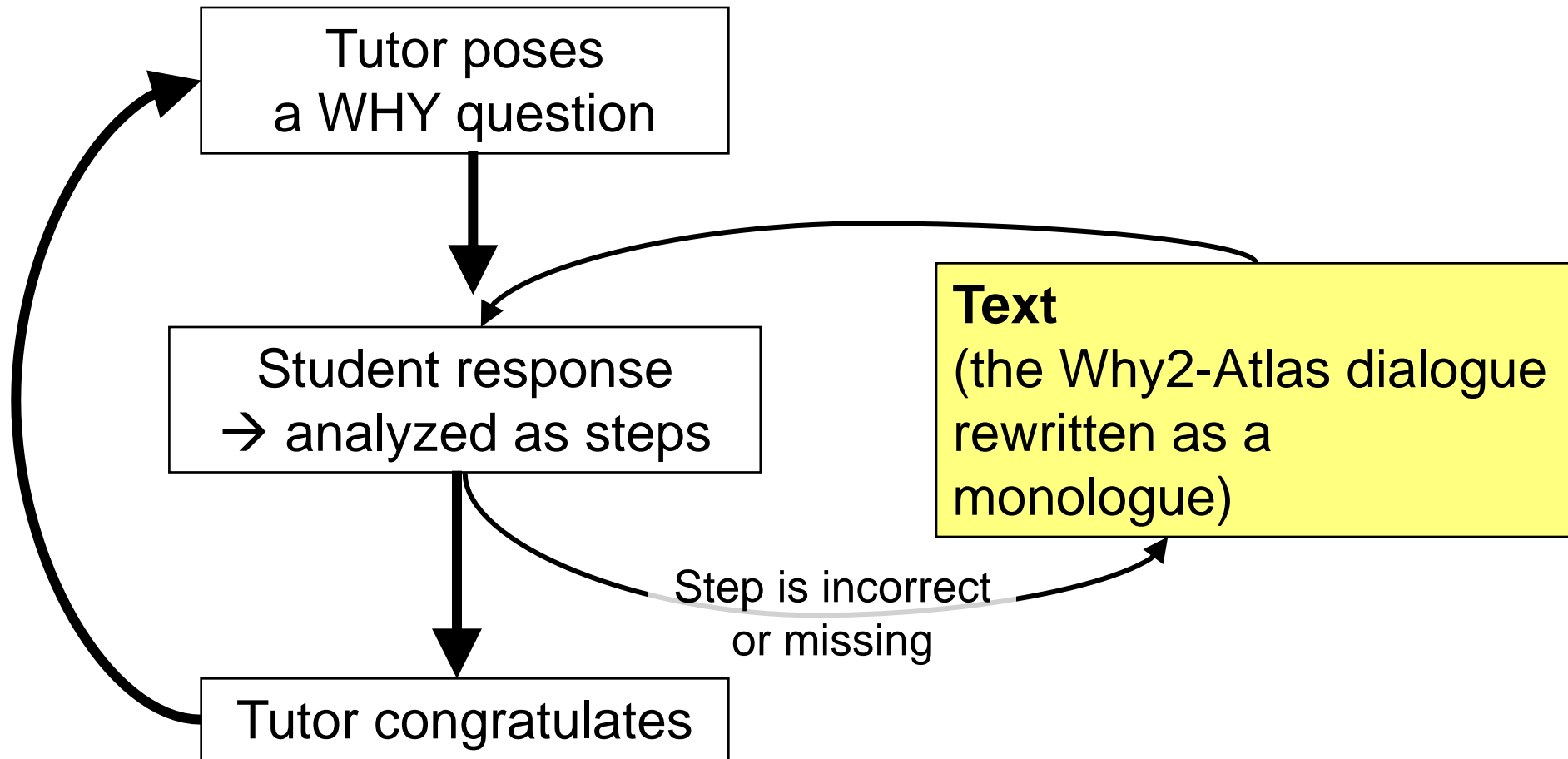
Why2-Atlas



Why2-AutoTutor

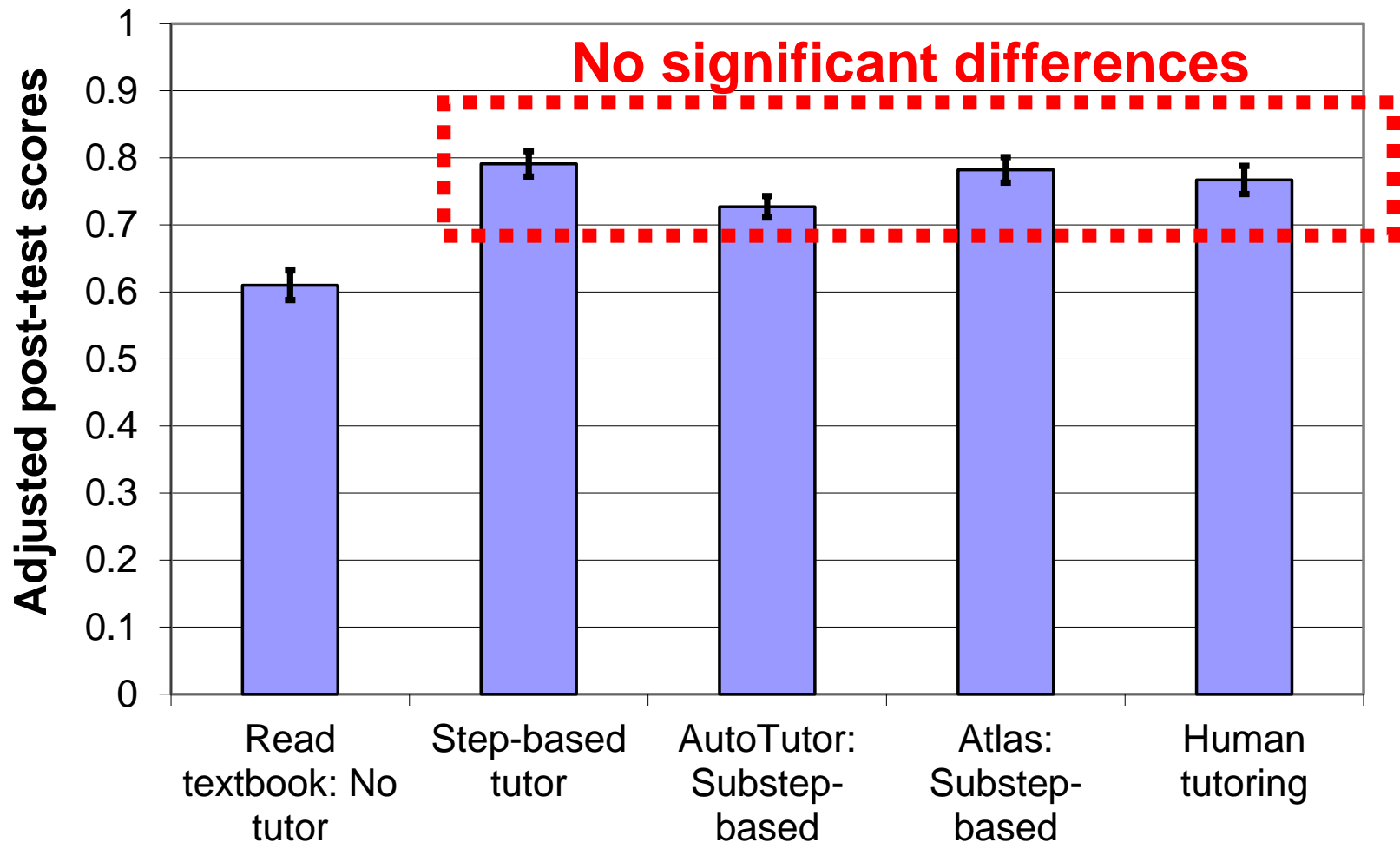


A step-based tutor: A text explanation with same content



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

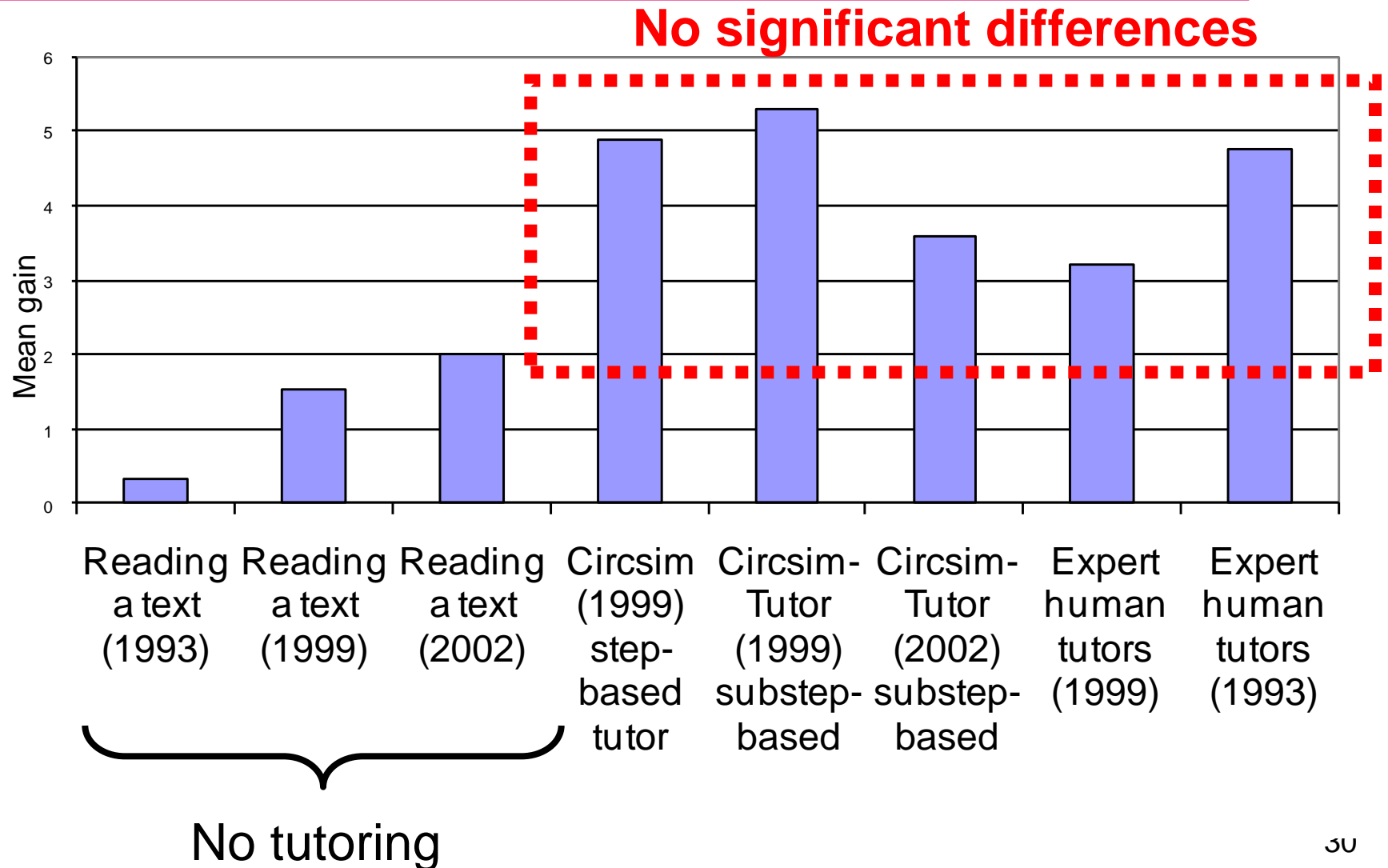
◆ Other studies

◆ Meta-analysis

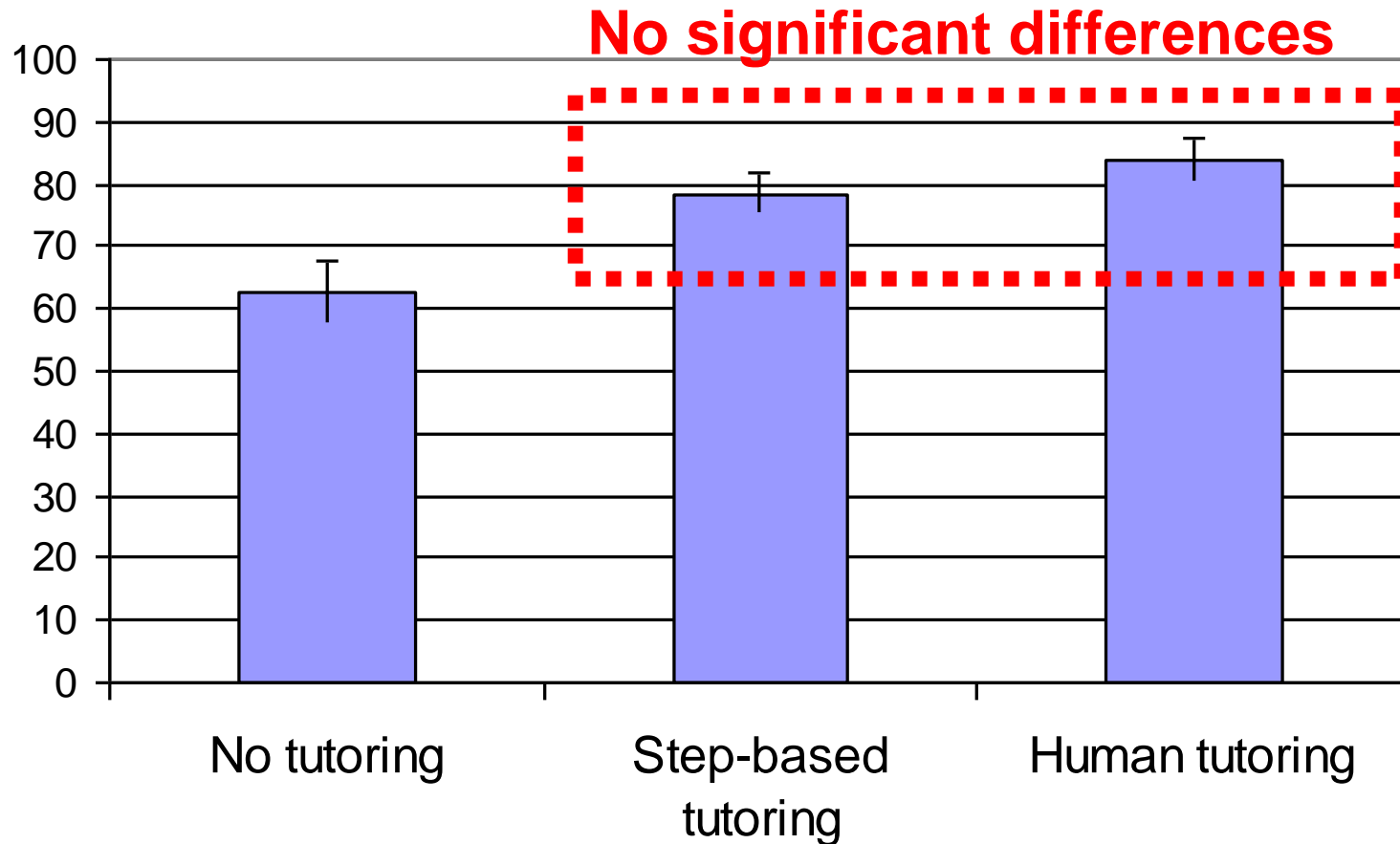


Next

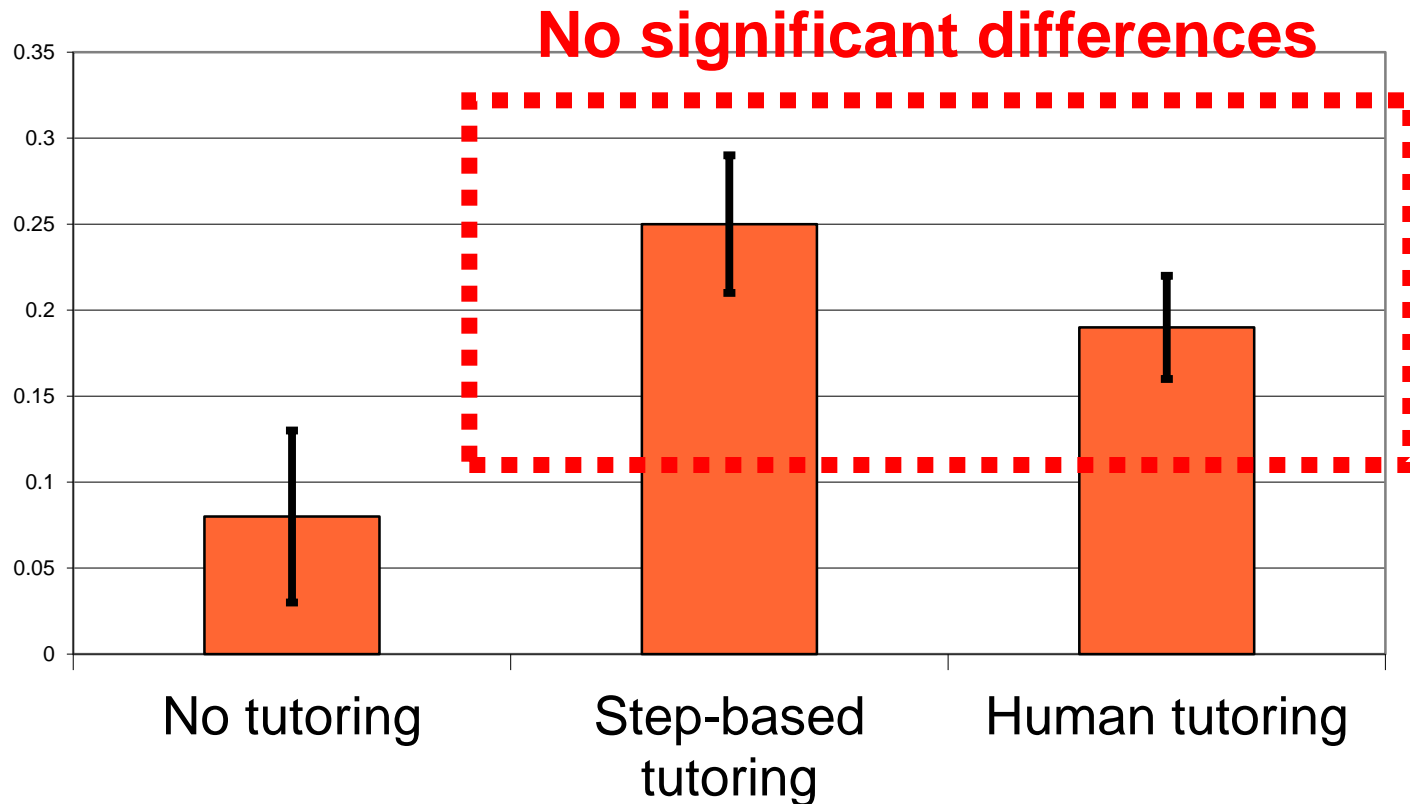
Evens & Michael (2006) also show
human tutoring = substep-based tutoring =
step-based tutoring



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



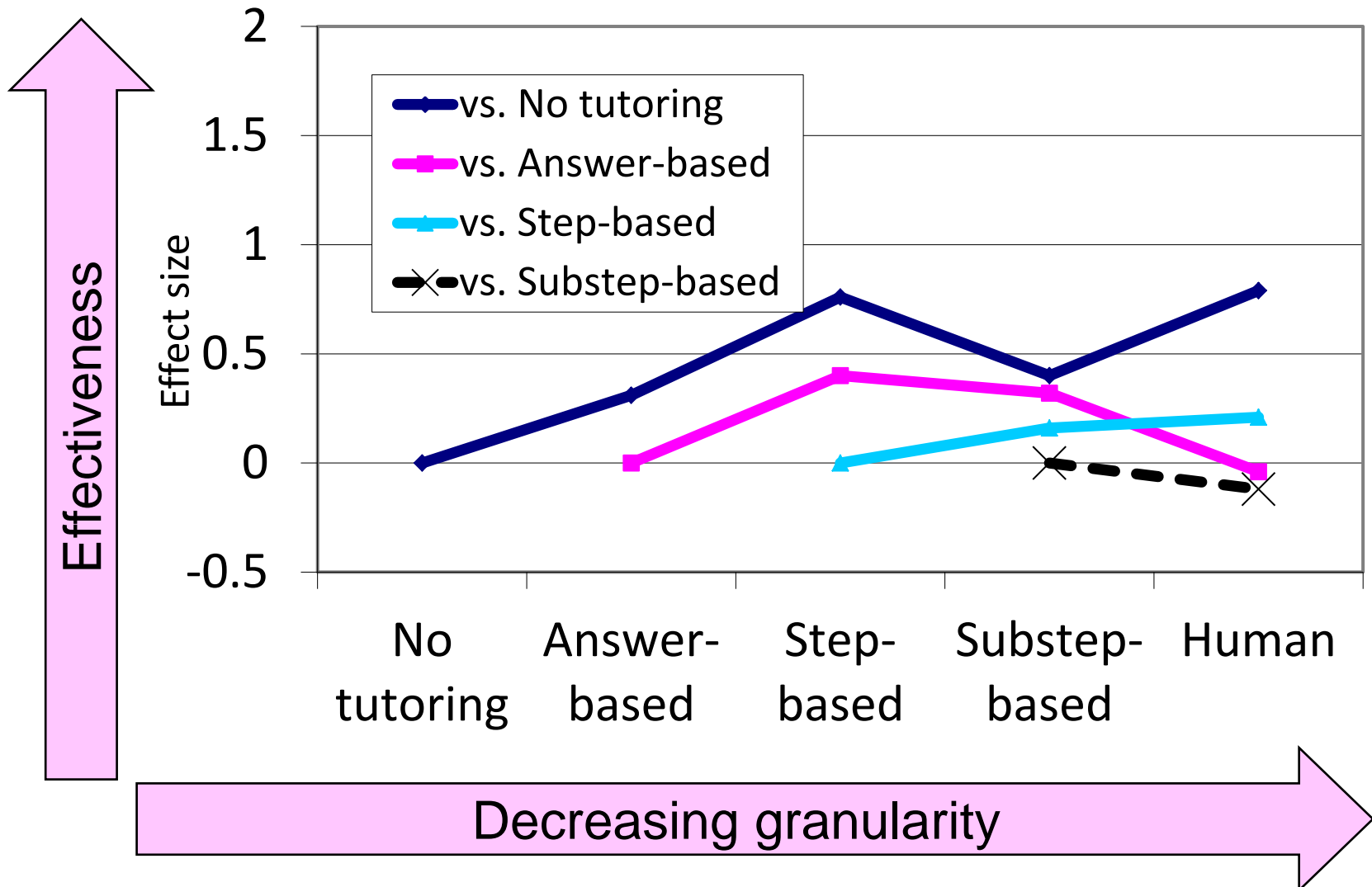
Next

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%

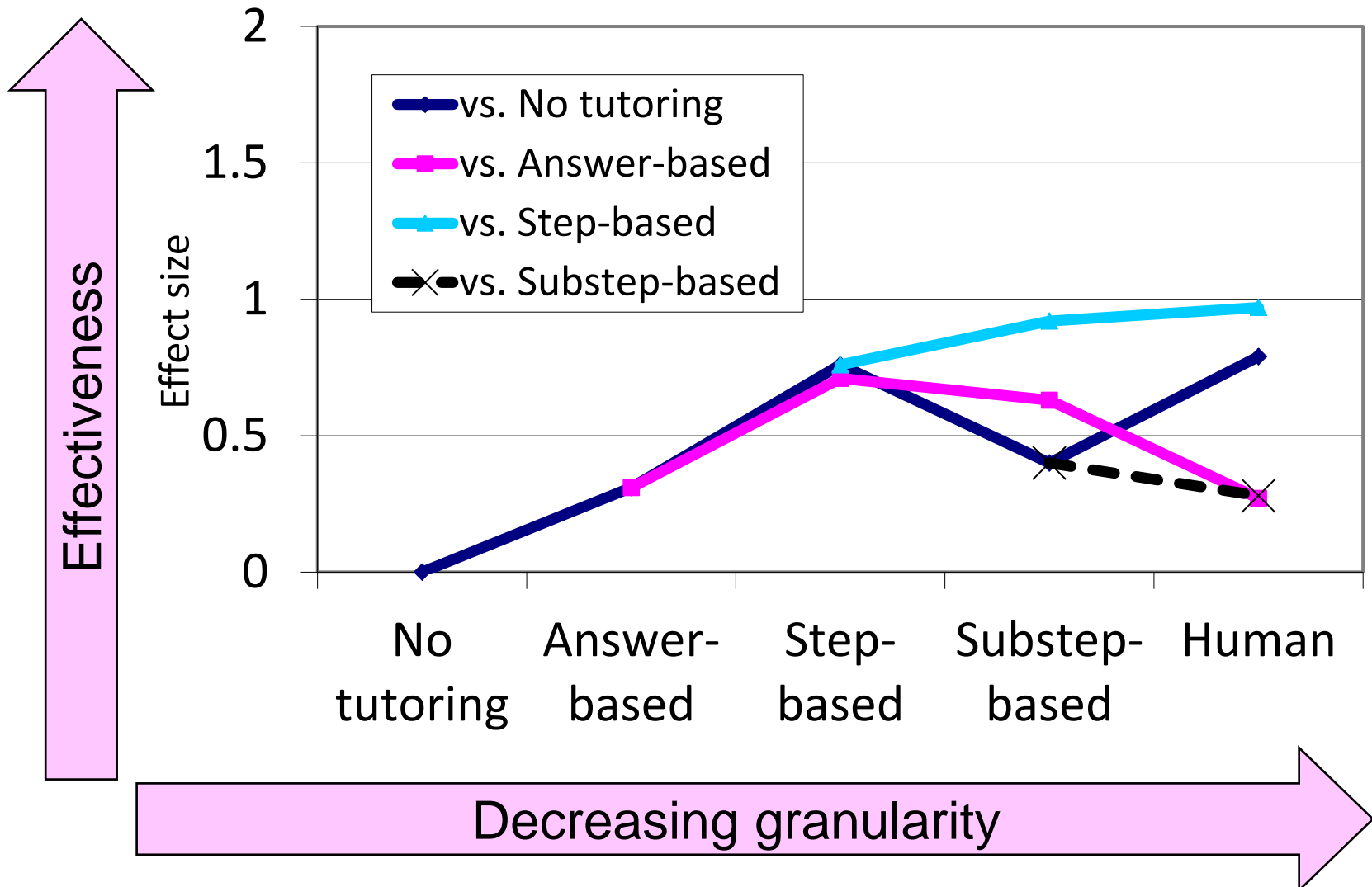
Graphing all 10 comparisons:

graph is hard to understand...

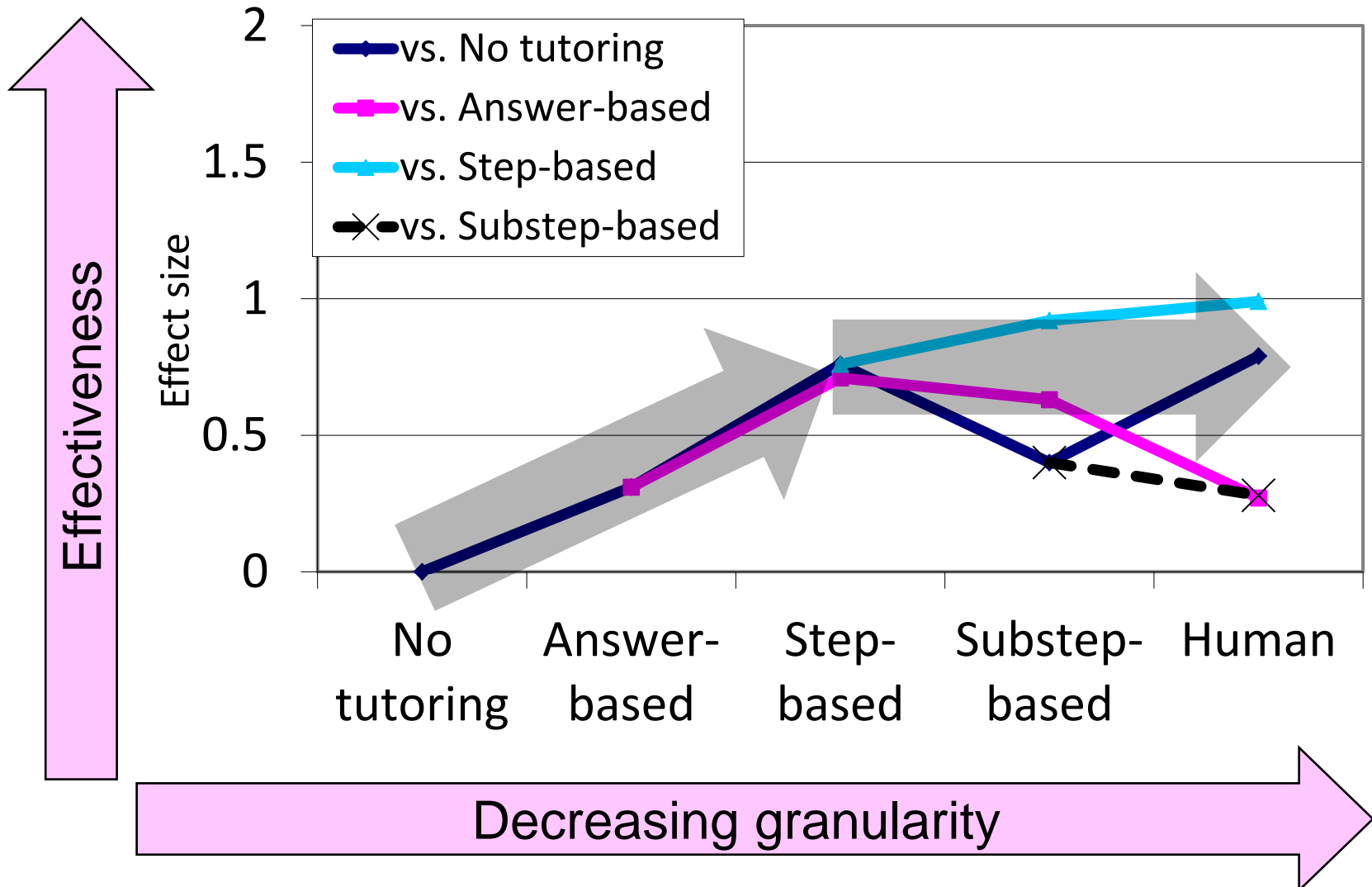


Graphing all 10 comparisons:

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)

Student utterance

The screenshot shows a window titled "Student Interface" with a blue title bar. It contains a "Problem Statement" tab, a "Dialog History" tab, and tabs for "Variables" and "Equations". The "Problem Statement" tab is active, displaying a physics problem about a rock's kinetic energy and work done. The "Dialog History" tab is also visible, showing a conversation between a student and a tutor. The "Variables" tab is active, displaying a list of variables and their values. The "Equations" tab is also visible.

Problem Statement

Dialog History

Variables

Equations

You said: definition of kinetic energy

Tutor said: Okay. Please write the equation for how the definition of kinetic energy applies to this problem at T1.

You said: value='ke1=1/2*m*v1^2'

Tutor said: Now it is easy to calculate the magnitude of v1. The magnitude of v1 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 T1 from the rock's kinetic energy at T1?

You said: no, we cannot

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

OK

Comments

Problem Statement

Dialog History

Variables

Equations

A 0.6 kg rock in space has a velocity of magnitude 2.0 m/s at point A and kinetic energy of 7.50 J

at point B. What is the net work done on the rock as it moves from A to B?

We define T0: the time point when the rock is at point A.

T1: the time point when the rock is at point B.

v0 The velocity of the rock during T0 is 2.0 m/s at an unknown orientation

KE0 The kinetic energy of the rock at T0 is 1.20 J

v1 The velocity of the rock during T1 is 5.0 m/s at an unknown orientation

KE1 The kinetic energy of the rock at T1 is 7.50 J

Wnet01 The work done on the rock

TME0 The total mechanical energy of the system at T0

TME1 The total mechanical energy of the system at T1

Tutor feedback

Student enters an equation (step)

Tutor embeds conceptual

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

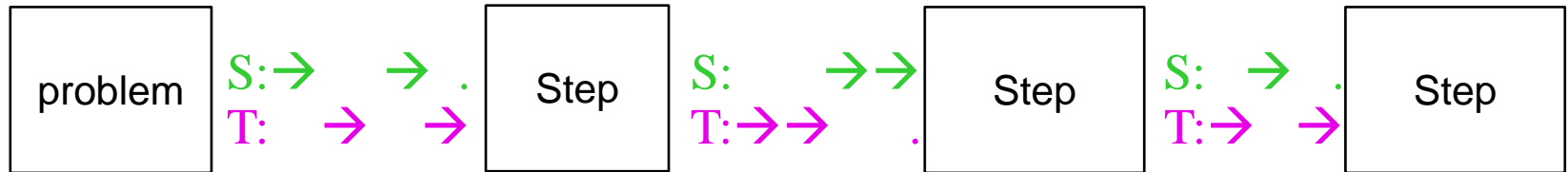
- ✓ Embedding conceptual in procedural
 - Cordillera produces better conceptual learning ($d=0.49$) than Andes, and Andes \approx human
- ◆ Machine learning of pedagogical tactics
- ◆ Meta-strategic scaffolding



Next

A self-improving tutoring system

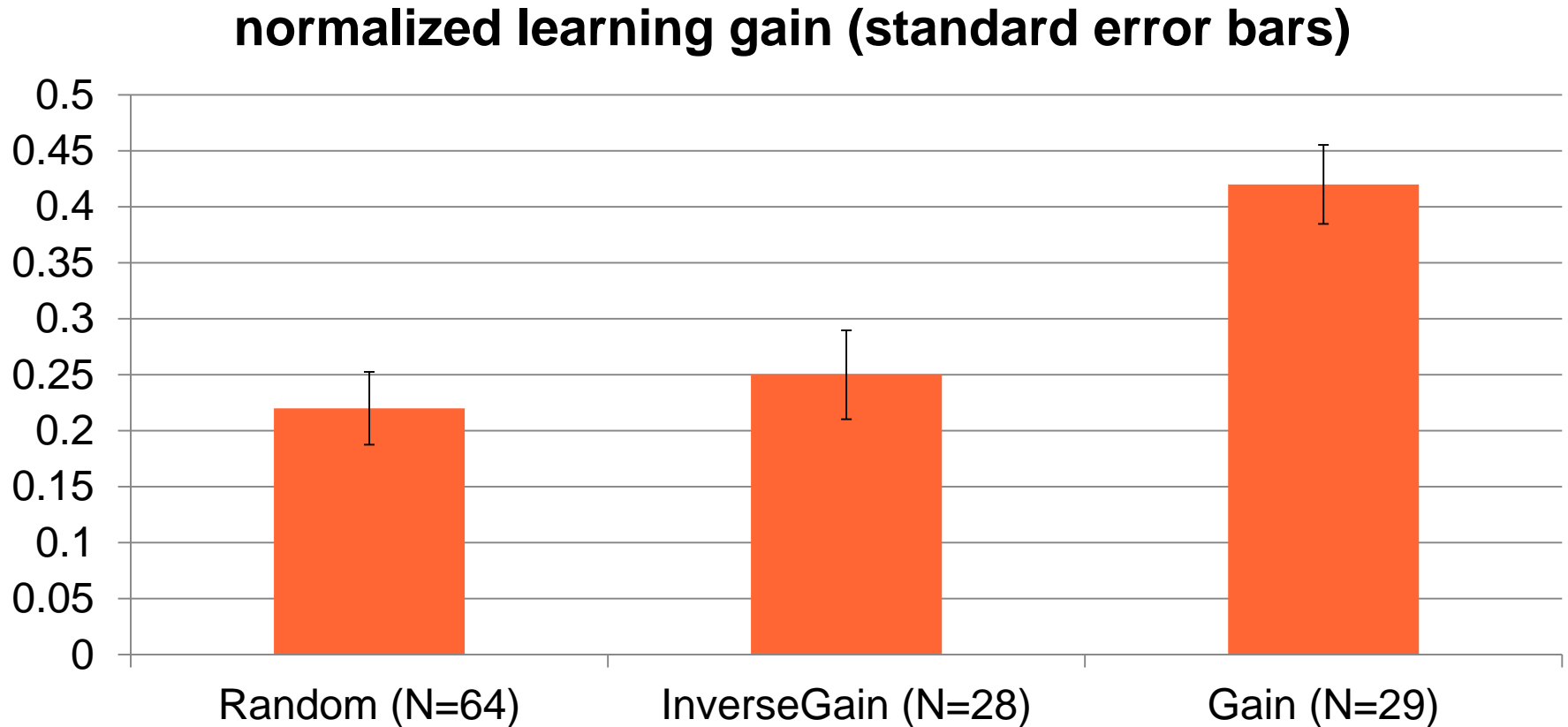
- ◆ Dialogue-based physics tutor (Cordillera)
- ◆ Chooses between *elicit* and *tell*



- ◆ 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 \approx 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 $V_x = V \cos(V_\theta)$...
- ◆ 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

A physics problem

The screenshot displays the Pyrenees user interface with four windows:

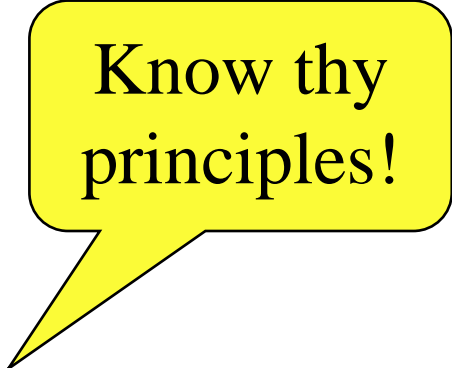
- Dialogue:** A text-based conversation between a student and a tutor. The student asks for help with a physics problem, and the tutor provides guidance through a series of questions and answers.
- Problem Statement:** A window containing a photograph of a cumulonimbus cloud and a text description of the physics problem: "Calculate the magnitude of the instantaneous velocity (speed) at which a hailstone, falling from 9000 meters out of a cumulonimbus cloud, would strike the ground. Assume the hailstone starts from rest and that air friction is negligible. Time 1 is when the hailstone..."
- Drawings:** A window showing a 2D coordinate system with a vertical y-axis and a horizontal x-axis. The origin is labeled '0'. The y-axis is labeled '+Y0' and the x-axis is labeled '+X0'. A point 'P' is marked on the y-axis, and a vector 'v' is shown pointing downwards from 'P'.
- Equations:** A window showing a list of equations and their derivations. The equations are:
 - 1) For $v_f y_0$:
 - 2) $v_f y_0 = -v_f$The text below the equations states: "For v_f : Projection of the velocity of the hailstone on the y-axis. The vector is opposite the axis"

Student-tutor
dialogue

Same 3 windows as
Andes, but read only

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



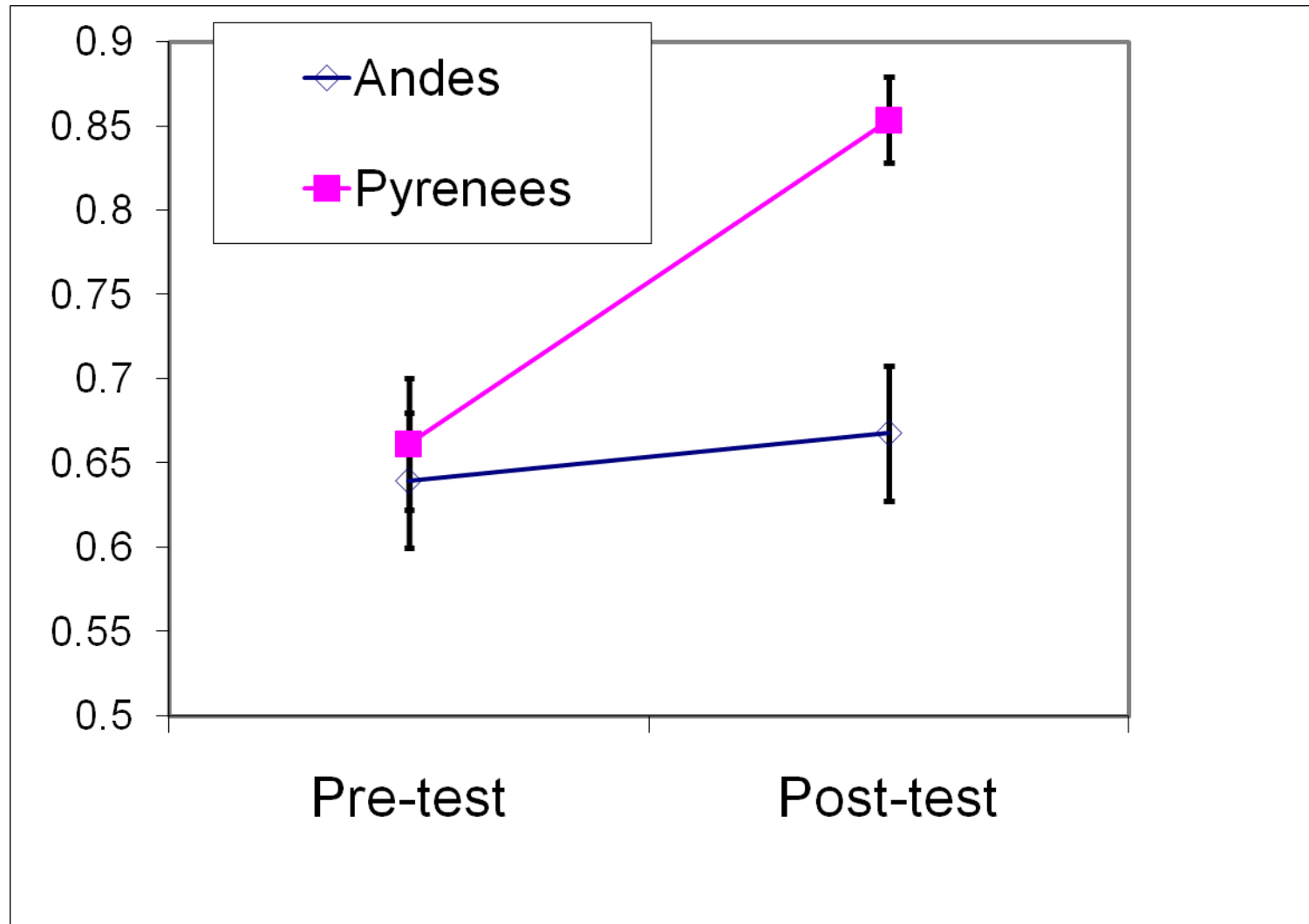
Know thy principles!

Experimental Procedure

	Pyrenees group	Andes group
Probability Instruction	Pyrenees	Andes
Physics Instruction	: Andes	

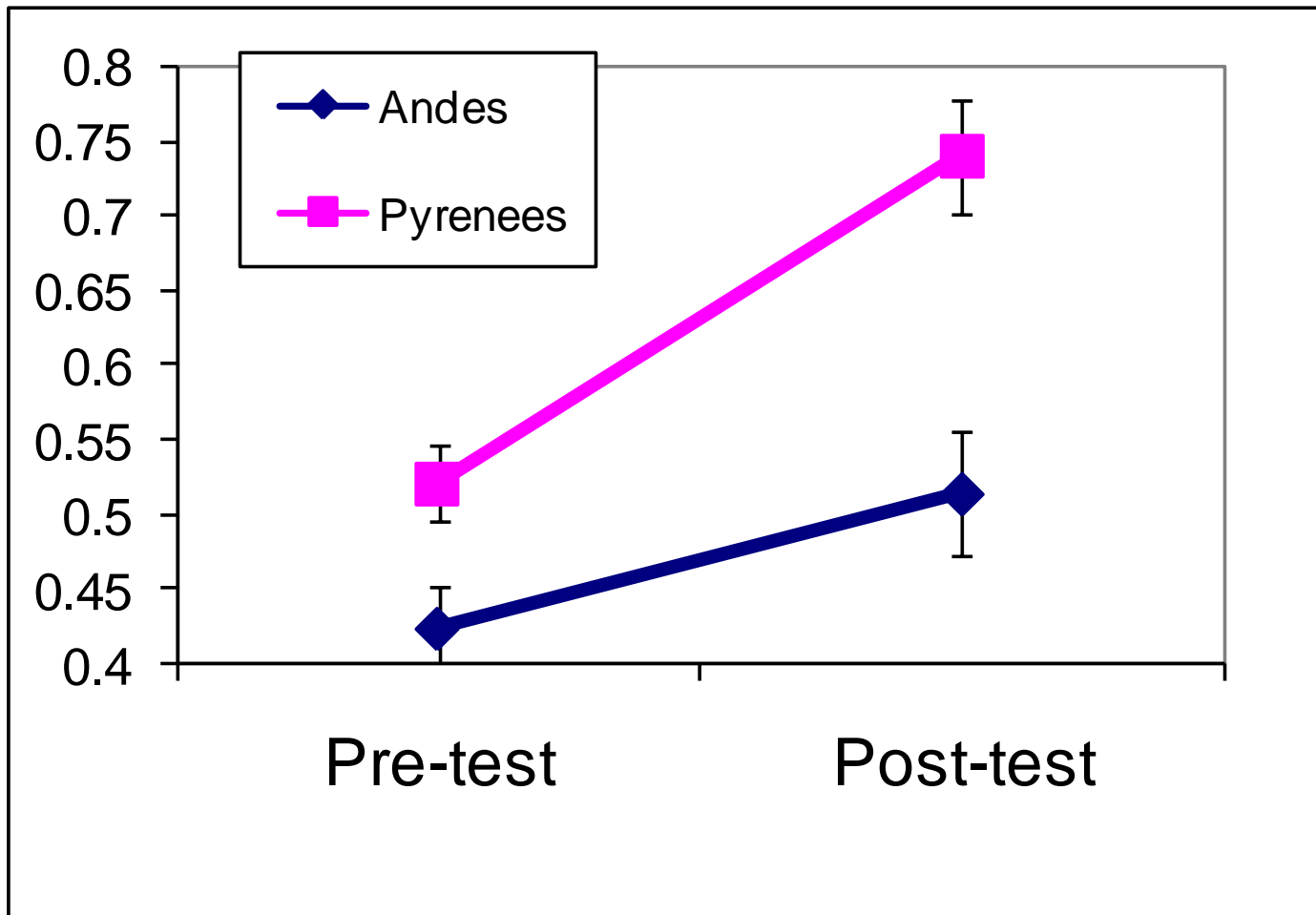
Instruction =
pre-training → pre-test → training → post-test

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 \approx 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.