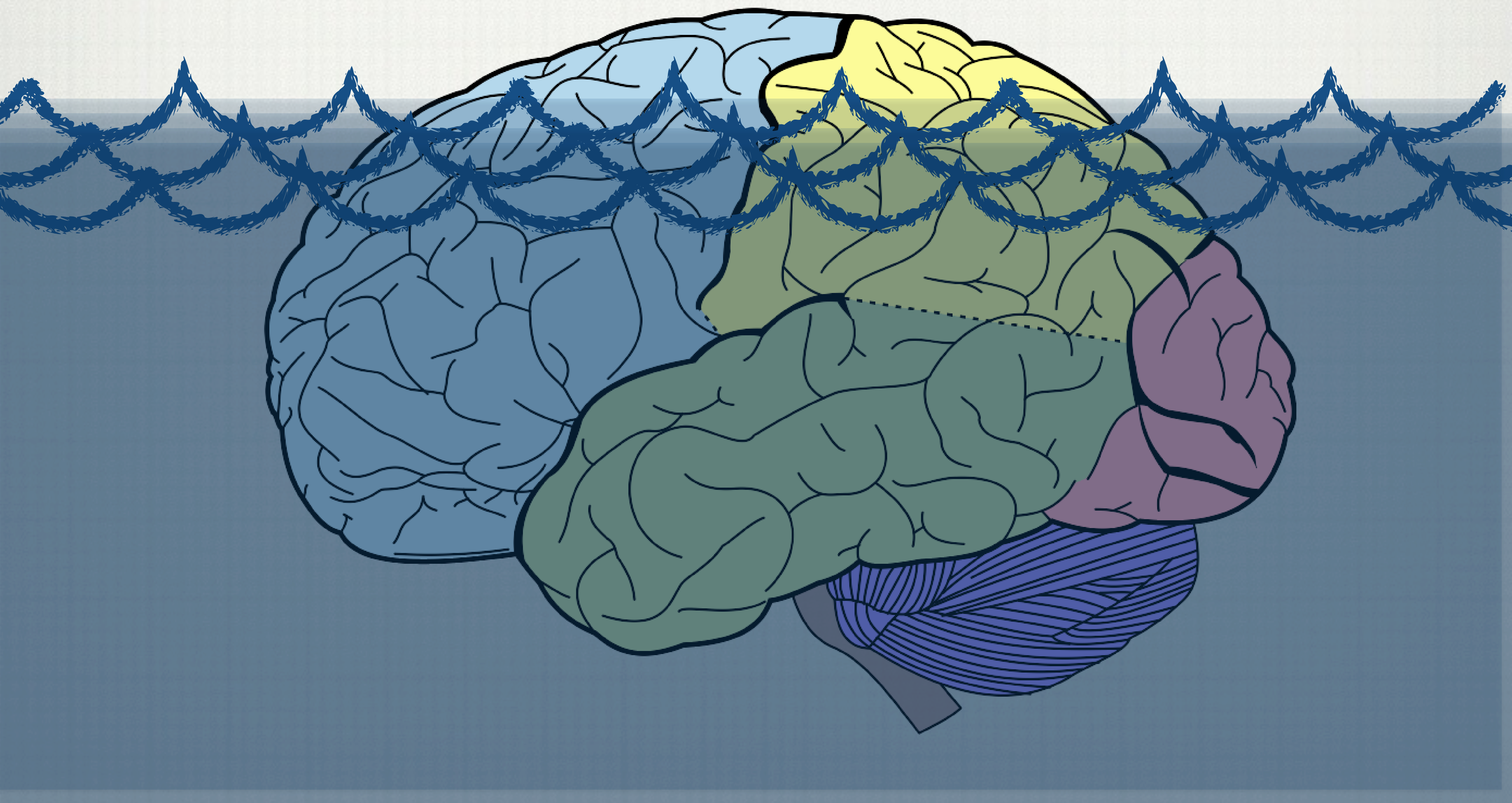
What can machine learning do for open education?

Geoff Gordon
CMU Machine Learning
ggordon@cs.cmu.edu

Civilization advances by extending the number of important operations which we can perform without thinking about them.

—Alfred North Whitehead, 1911



Contribution of ML

- Machine learning can help us understand how students learn

Contribution of ML

- Machine learning can help us understand how students learn
  - Not just any ML, but latent variable (“hidden feature”) discovery

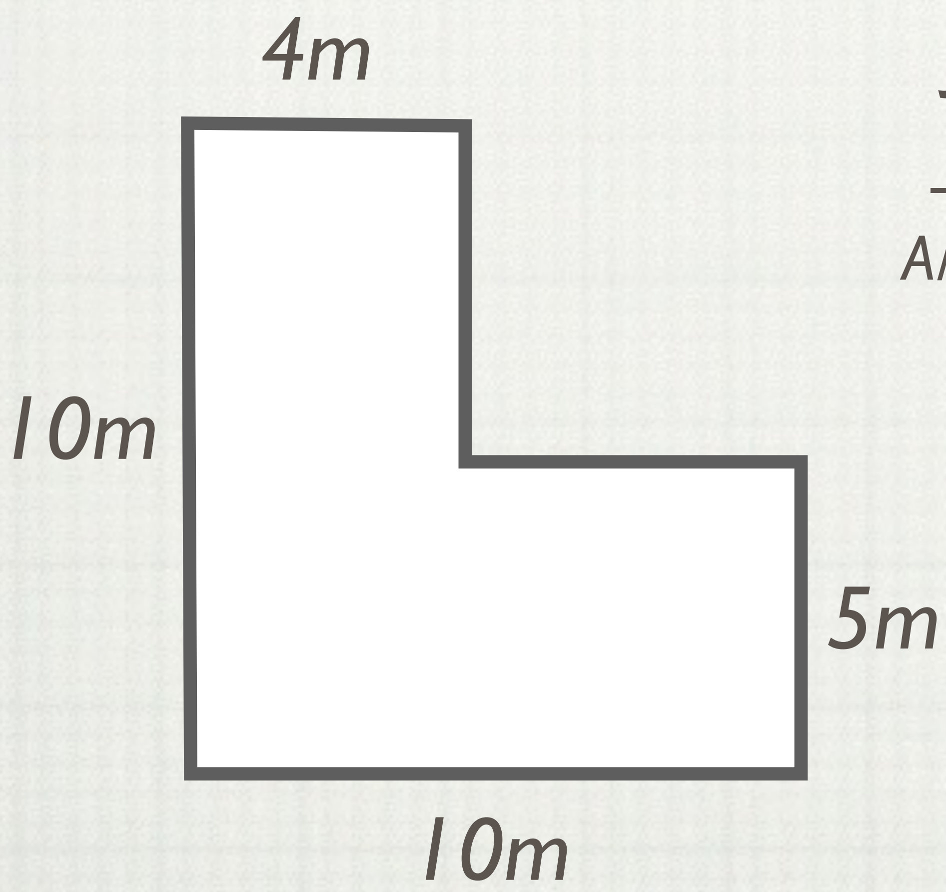
Contribution of ML

- Machine learning can help us understand how students learn
  - Not just any ML, but latent variable ("hidden feature") discovery
    - Not just any latent variables, but highly structured ones

Why bother?

- Student feedback
  - what does the student know?
  - what are common causes for the mistake the student just made?
- Instructor feedback
  - what do the students know?
  - what skills does this course content address?
  - what skills **doesn't** this course content address?
- Evaluation
  - help design rubrics for (peer, instructor) grading
  - cluster submissions by similar approach, skill level, …
- Etc…

Goal: understand how students learn something

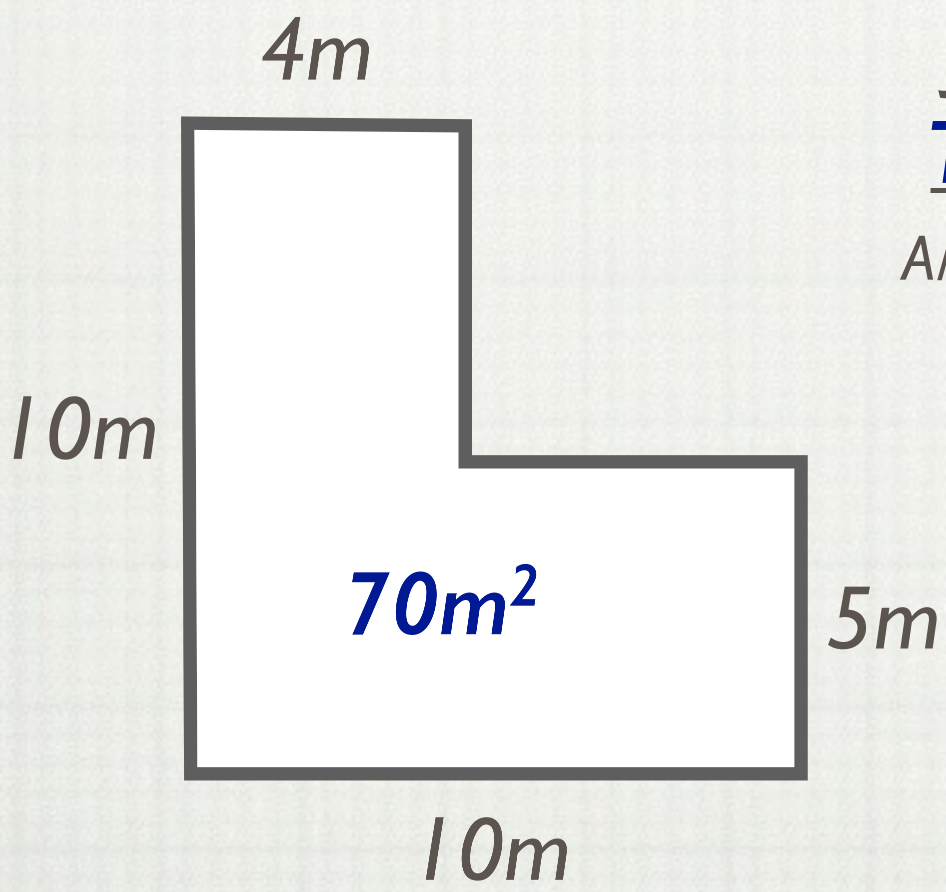


John took Joe for a ride on his boat.
___ boat was blue with a red stripe.

A/The/[]

$$3x + 4 = x + 10$$

Goal: understand how students learn something



John took Joe for a ride on his boat.
<u>The</u> *boat was blue with a red stripe.*

A/The/[]

$$3x + 4 = x + 10$$

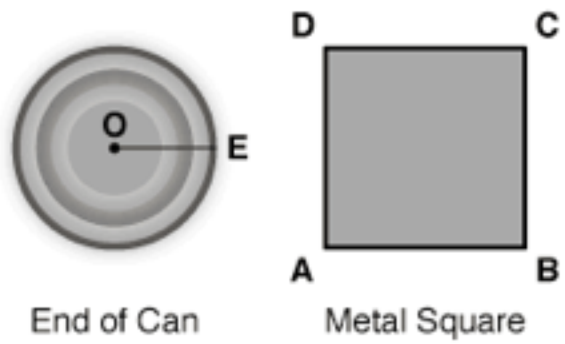
$$\mathbf{x = 3}$$

Ex: geometry tutor

Scenario

File Edit Tutor Windows Help

scenario



To make metal cans, the ends for the cans are stamped out of square pieces of metal. The part of the square that is left over is then recycled as scrap. The manufacturer needs to know the area of the scrap for each end. Then the total weight of the scrap can be figured out.

1. The can end has a radius of 4 inches. If an end is punched out of a square piece of metal measuring 8 inches on a side, find the square inches of the scrap.

2. The can end has a radius of 8 inches. If an end is punched out of a square piece of metal measuring 16 inches on a side, find the square inches of the scrap.

3. The can end has a radius of 12 inches. If an end is punched out of a square piece of metal measuring 24 inches per side, find the square inches of the scrap.

NOTE: To find the area of the scrap metal remaining, you might have to first find the area of the can end, and the area of the metal square

For this problem use an approximate value for pi. $\pi \approx 3.14$

Problem Making Cans

Worksheet

File Edit Tutor Worksheet Windows Help

worksheet

	radius of the end of the can	length of the square ABCD	Area of the scrap metal	AREA OF SQUARE ABCD	AREA OF END OF CAN
Unit	inches	inches	square inches	SQUARE INCHES	SQUARE INCHES
Diagram Label		AB			
Question 1	4	8	13.76	64	50.24
Question 2	8	16	55.04	256	200.96
Question 3	12	24	123.84	576	452.16

Spreadsheet Calculation OFF

_Geo Unit01-6's skills

File Edit Windows

skills

Adding/subtracting areas

1 Area / 6 Composite / Making Cans

http://www.carnegielearning.com/

STEP-LEVEL DATA

	radius of the end of the can	length of the square ABCD	Area of the scrap metal	AREA OF SQUARE ABCD	AREA OF END OF CAN
Unit	inches	inches	square inches	SQUARE INCHES	SQUARE INCHES
Diagram Label		AB			
Question 1	4	8	13.76	64	50.24
Question 2	8	16	55.04	256	200.96
Question 3	12	24	123.84	576	452.16

one step = fill in a box

- Record right/wrong, timing, use of hints, …

Simplest model: Rasch / 1-parameter Item Response Theory

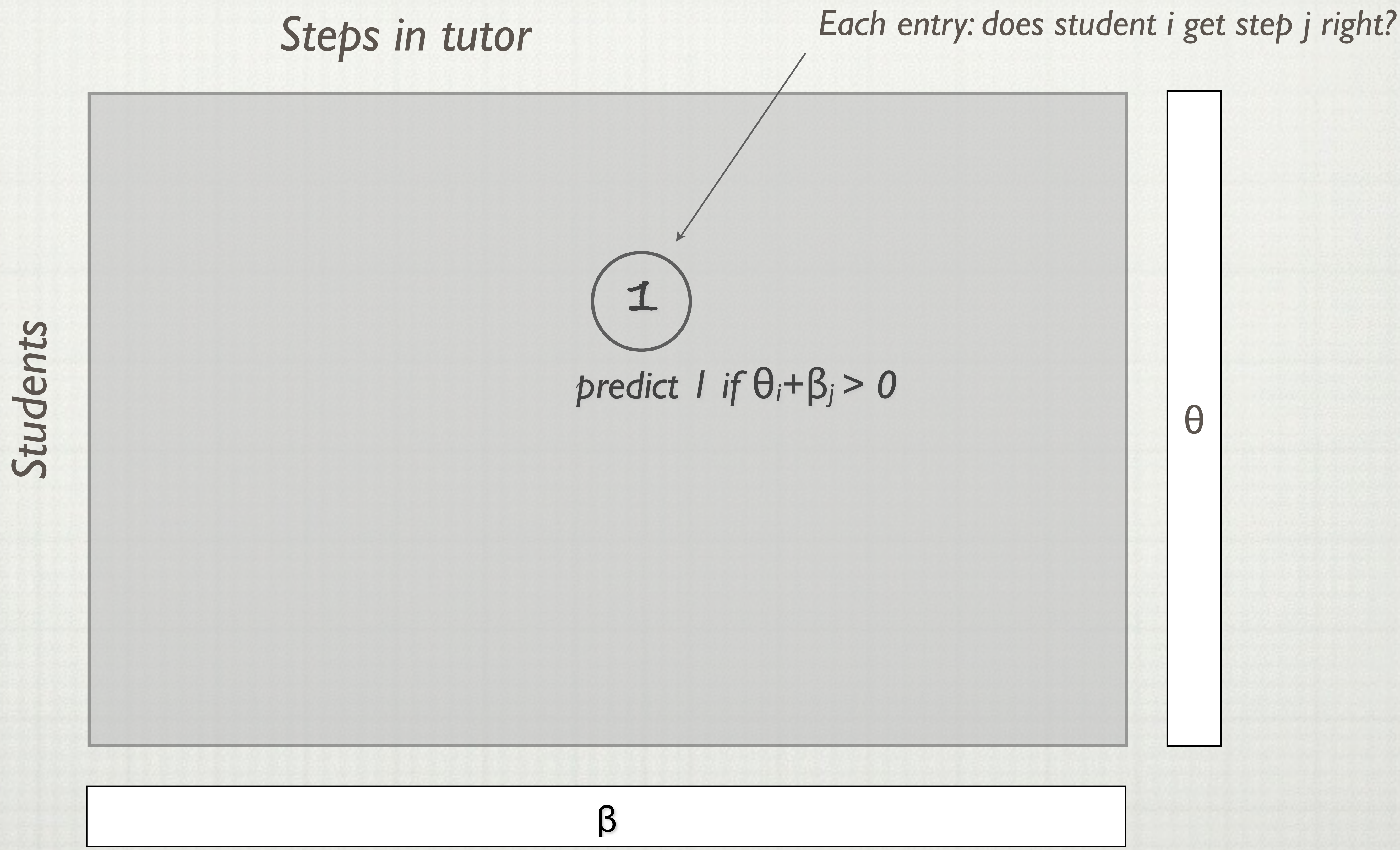
$$\ln\left(\frac{p_t}{1-p_t}\right) = \theta_{i_t} + \beta_{j_t}$$

p_t = *P(correct answer)*

i_t = *student ID* j_t = *step ID*

θ = *student mean (knowledge level)*
β = *item mean (easy/difficult)*

Simplest model: Rasch / 1-parameter Item Response Theory

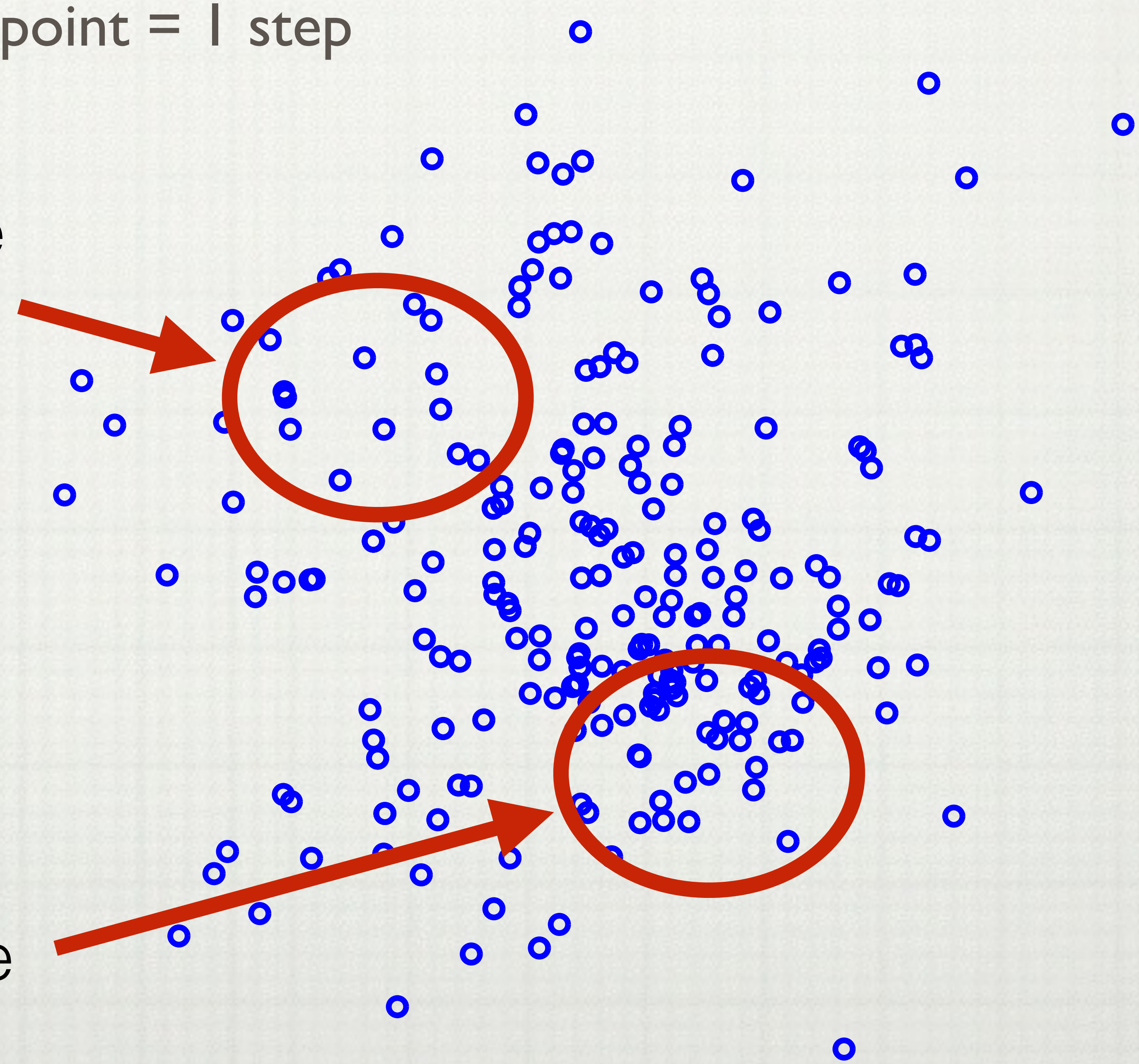


STRUCTURE: SIMILARITY AMONG STEPS

- Learn a "step map": each point = 1 step

Steps over here are more similar to each other…

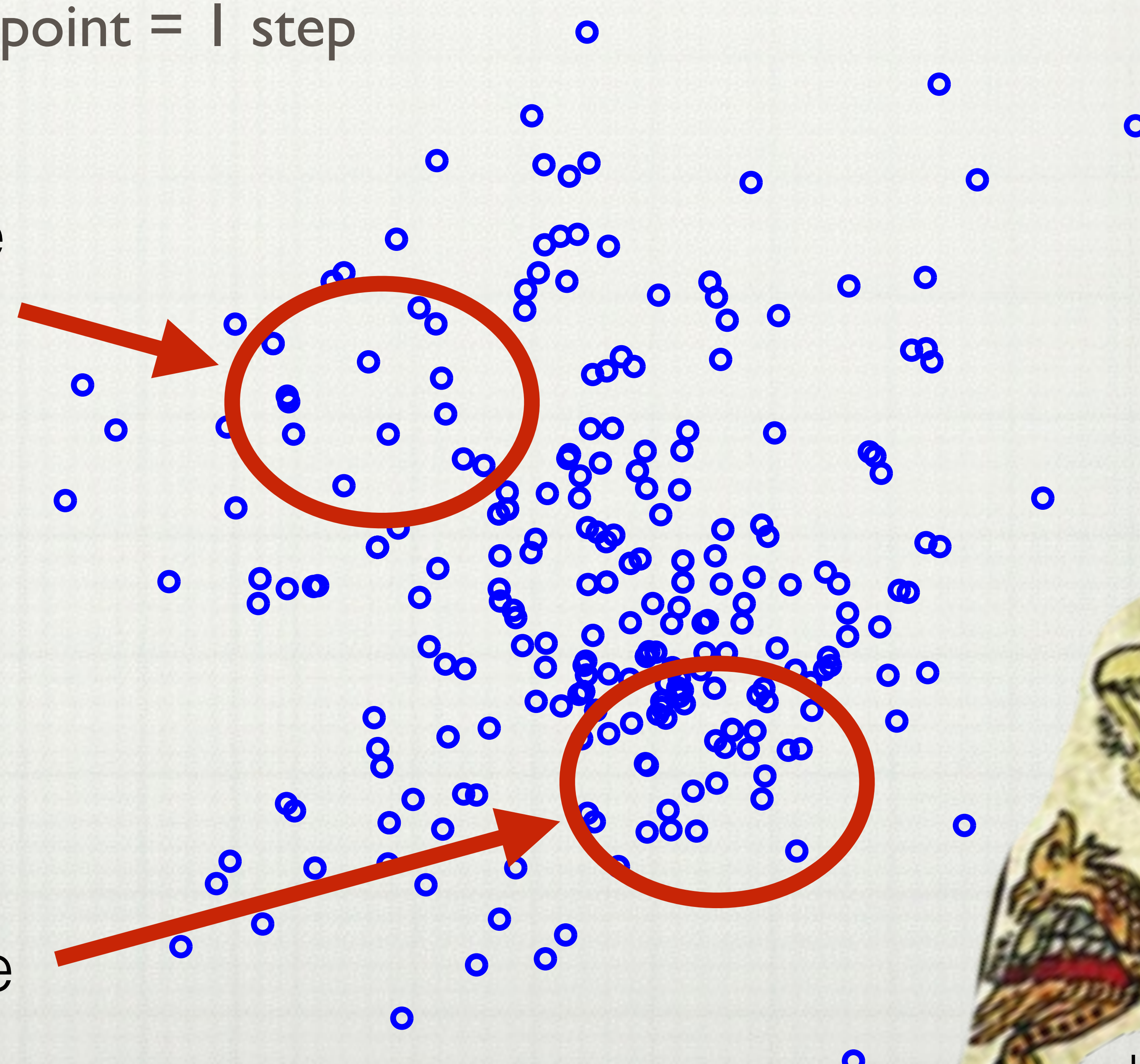
… than to steps over here



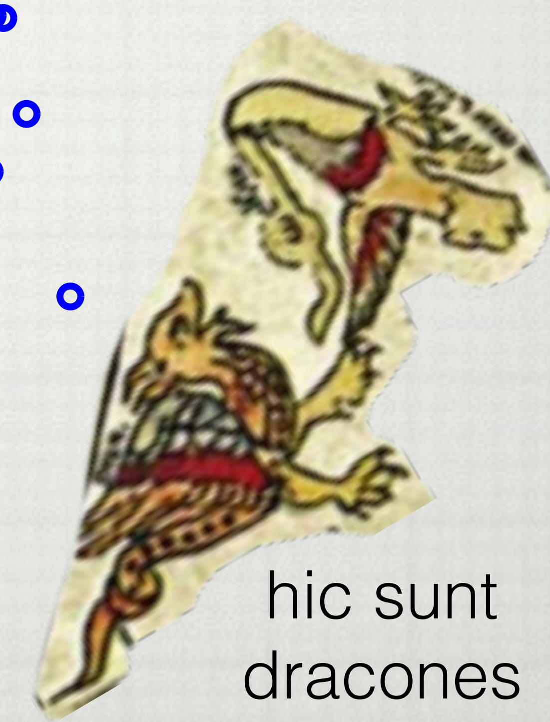
STRUCTURE: SIMILARITY AMONG STEPS

- Learn a "step map": each point = 1 step

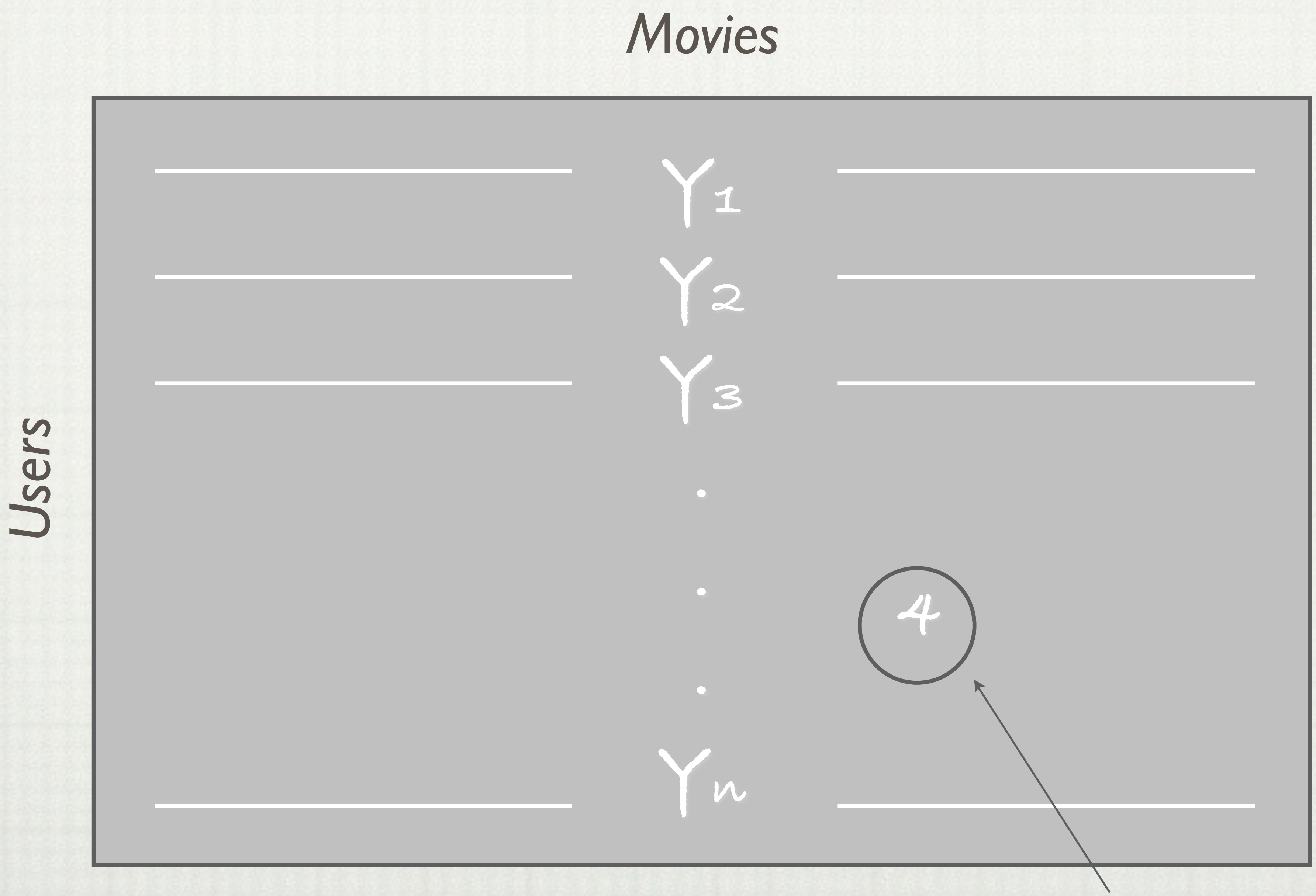
Steps over here are more similar to each other…



… than to steps over here

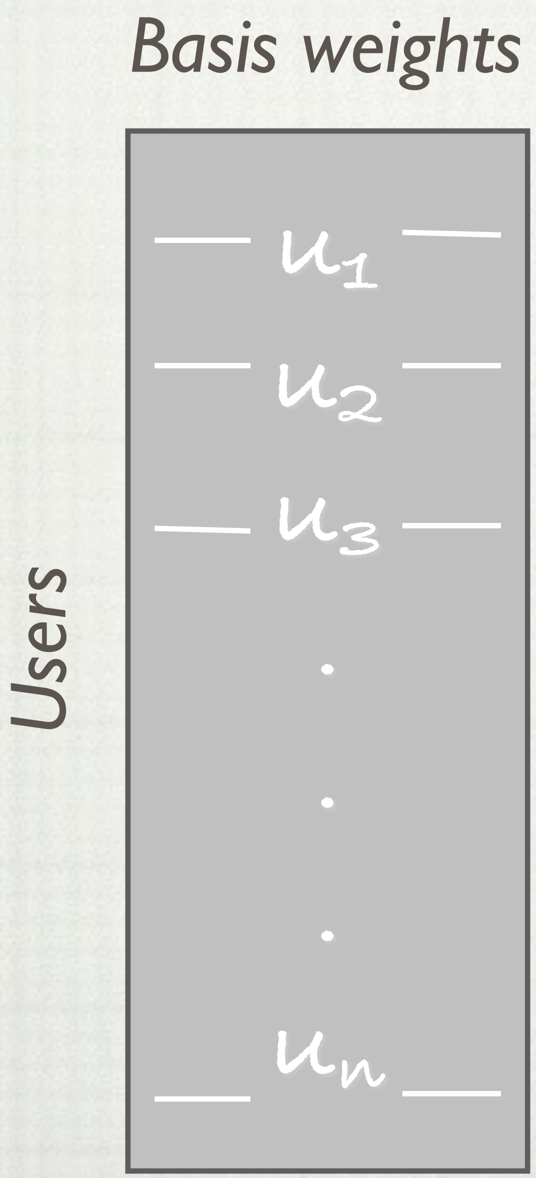


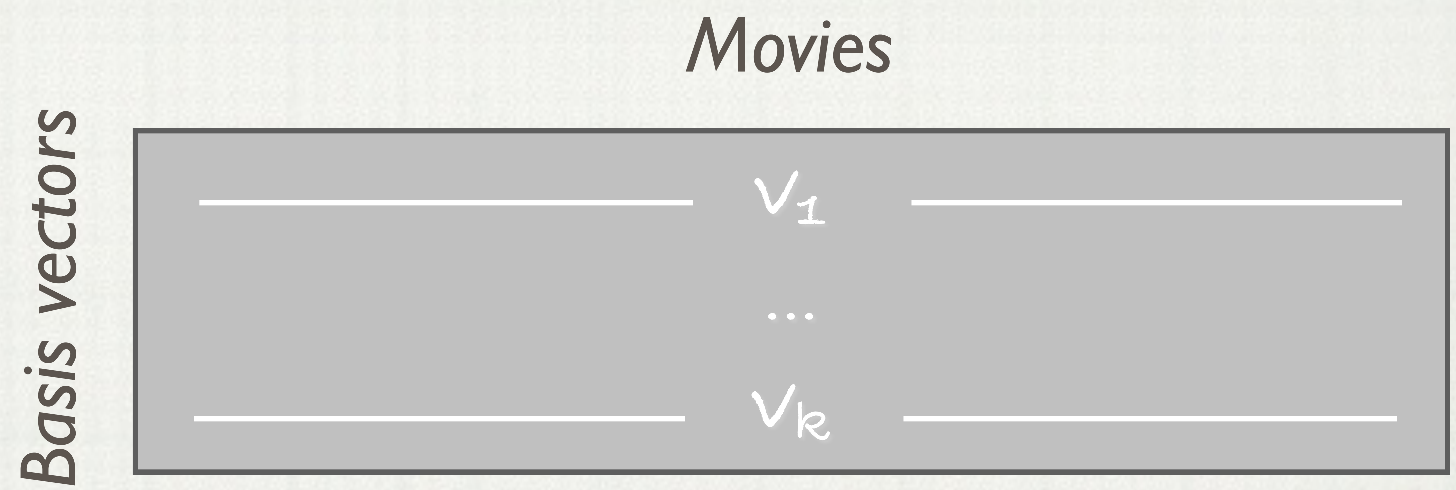
How principal components anaysis got famous



Each entry: how many stars does user i give to movie j?

Result of factoring

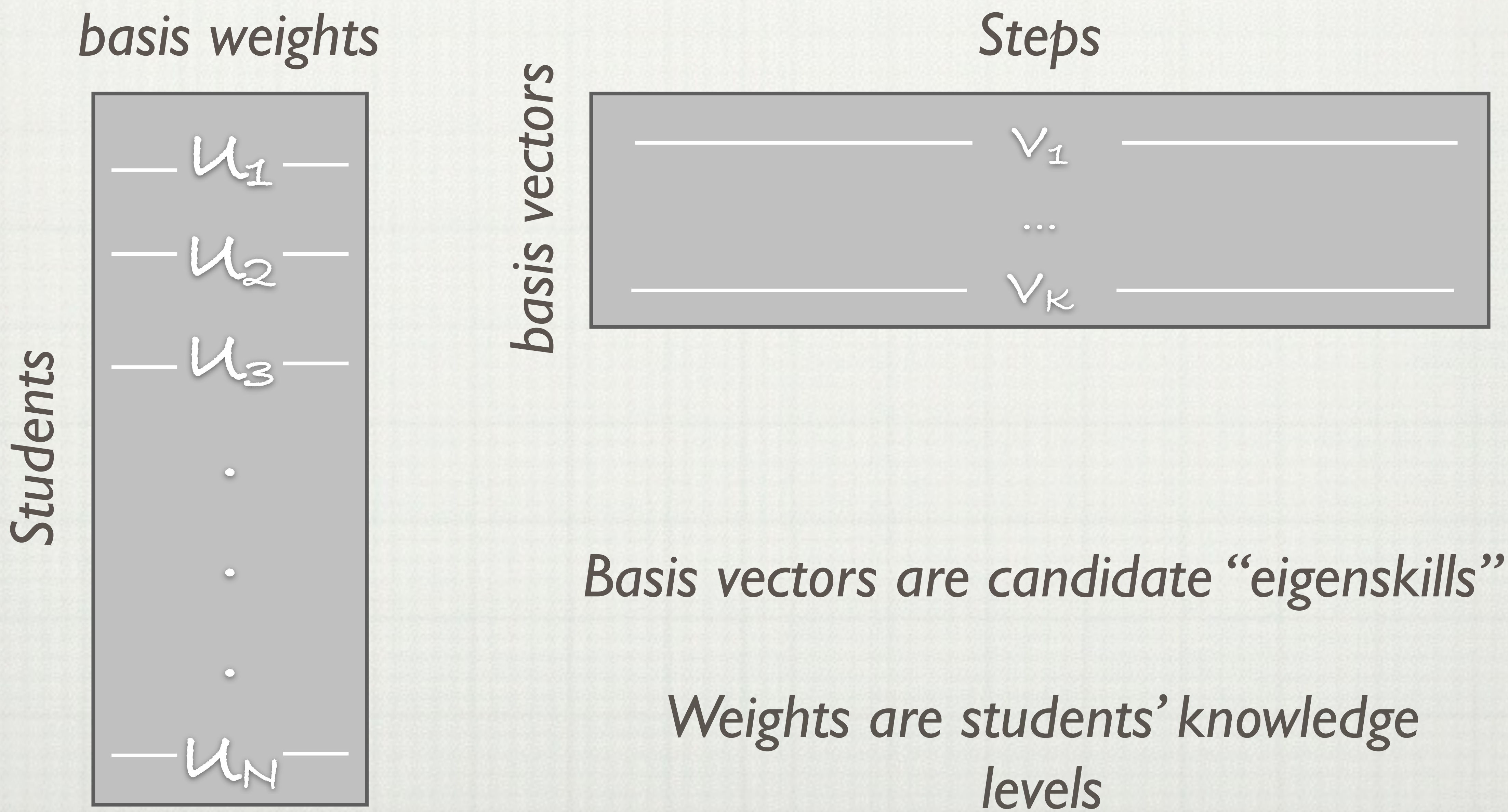




Low-d basis = latent variables

Basis vectors represent latent properties of movies, e.g., "is a comedy"

IN OUR CASE (STUDENT-STEP DATA)



Basis vectors are candidate "eigenskills"

Weights are students' knowledge levels

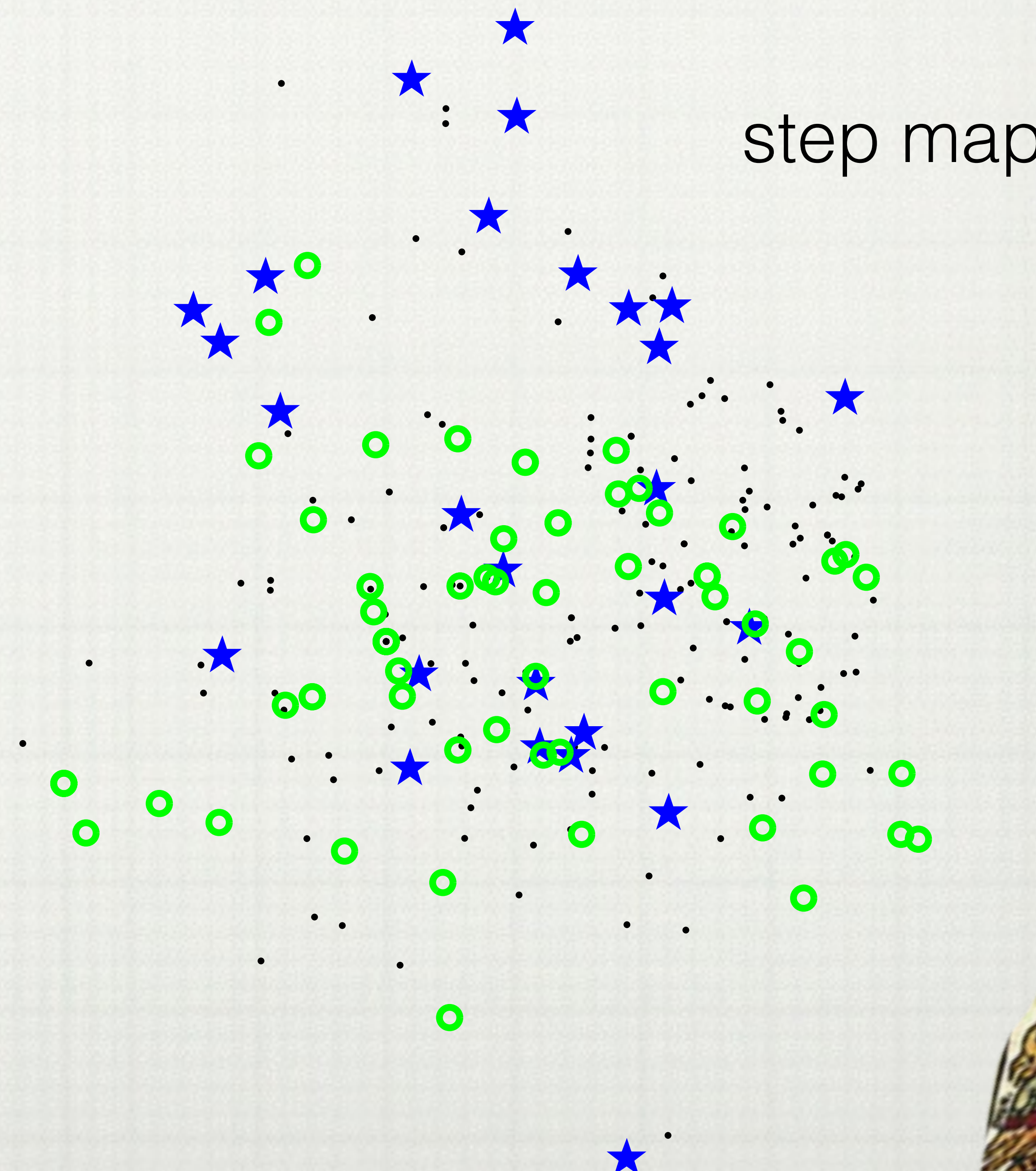
DOES IT WORK?

★ steps about pentagons

○ steps about circles

· other steps

Learned features let us predict held-out data better than chance ($\rho = .3$, $p < 0.0001$)

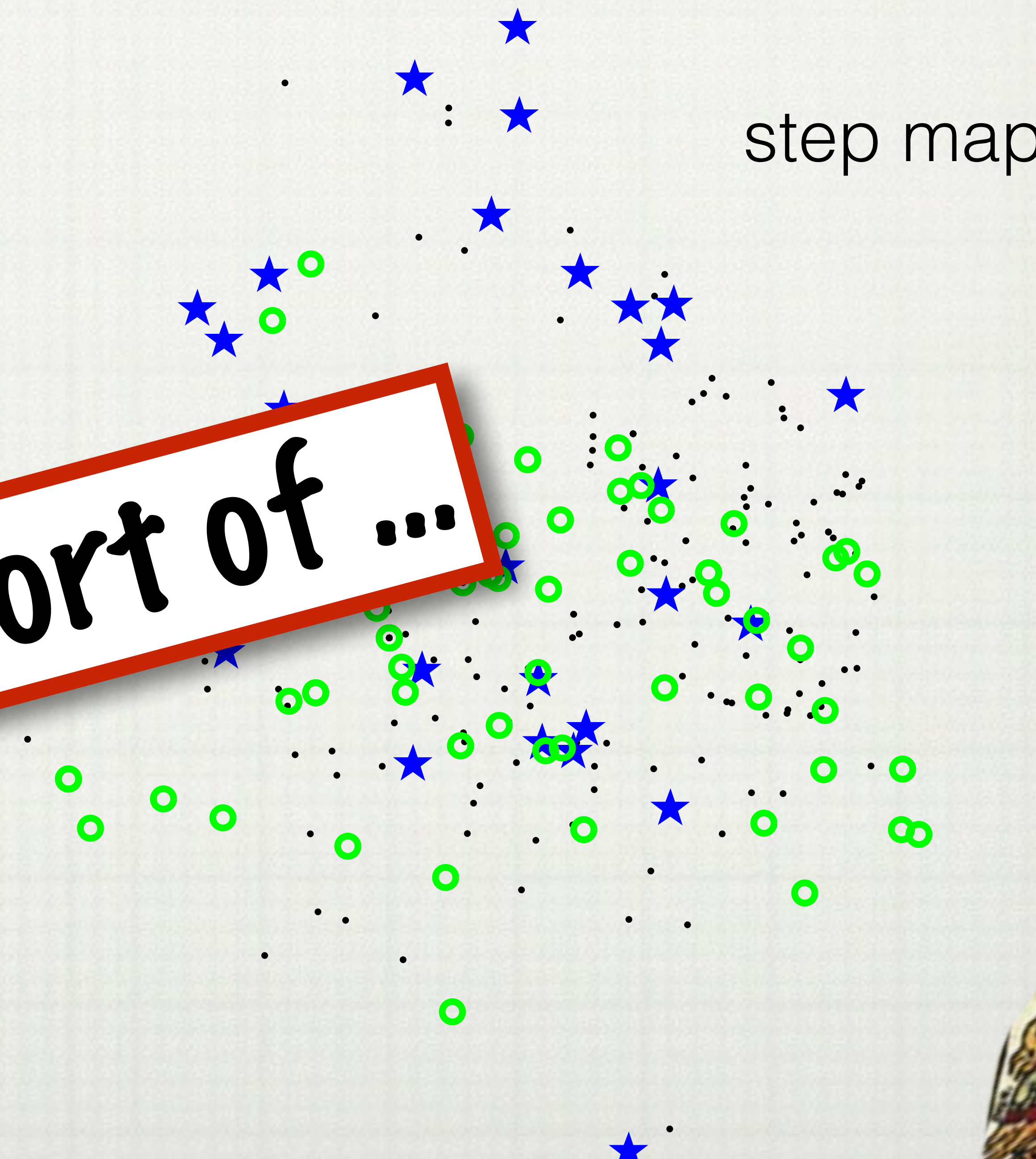


Does it work?

★ steps about pentagons

○ steps about circles

· other steps



Yes, sort of ...

Learned features le
predict held-out data b
than chance
($\rho = .3$, $p < 0.0001$)

STRUCTURE: PRACTICE MAKES PERFECT

- PCA ignores step order — clearly wrong…
- Add model of student learning to PCA
  - based on "additive factor model" [Draney et al., 1995]

STRUCTURE: PRACTICE MAKES PERFECT

- PCA ignores step order — clearly wrong…
- Add model of student learning to PCA
  - based on "additive factor model" [Draney et al., 1995]
- Result: predictions of held-out data get slightly better
  - $\rho = .45$ ($p < 0.01$ vs. plain PCA)
- Step map still looks the same
- Meh…

What we really want

- To be understandable to us humans, latents need to be sparse and binary (‘is about circles’, ‘requires subtracting areas’)
- Can’t do this fully automatically from this small data set (only 59 students, 370 steps)
- Challenge: can we discover sparse, binary, understandable latents automatically from MOOC-scale data?

"KC Hypothesis"

- Knowledge comes in atomic units ("KCs")
- Each KC is learned independently (no transfer)
  - transfer *among steps* mediated by common KCs
  - or prerequisite structure (can't learn algebra w/o knowing arithmetic)
- Each student has a (latent, scalar) proficiency level for each KC
  - learn/forget = transition to a higher/lower proficiency level
- Learning a KC happens only through exposure to that KC
  - problem, worked example, lecture, real life, …

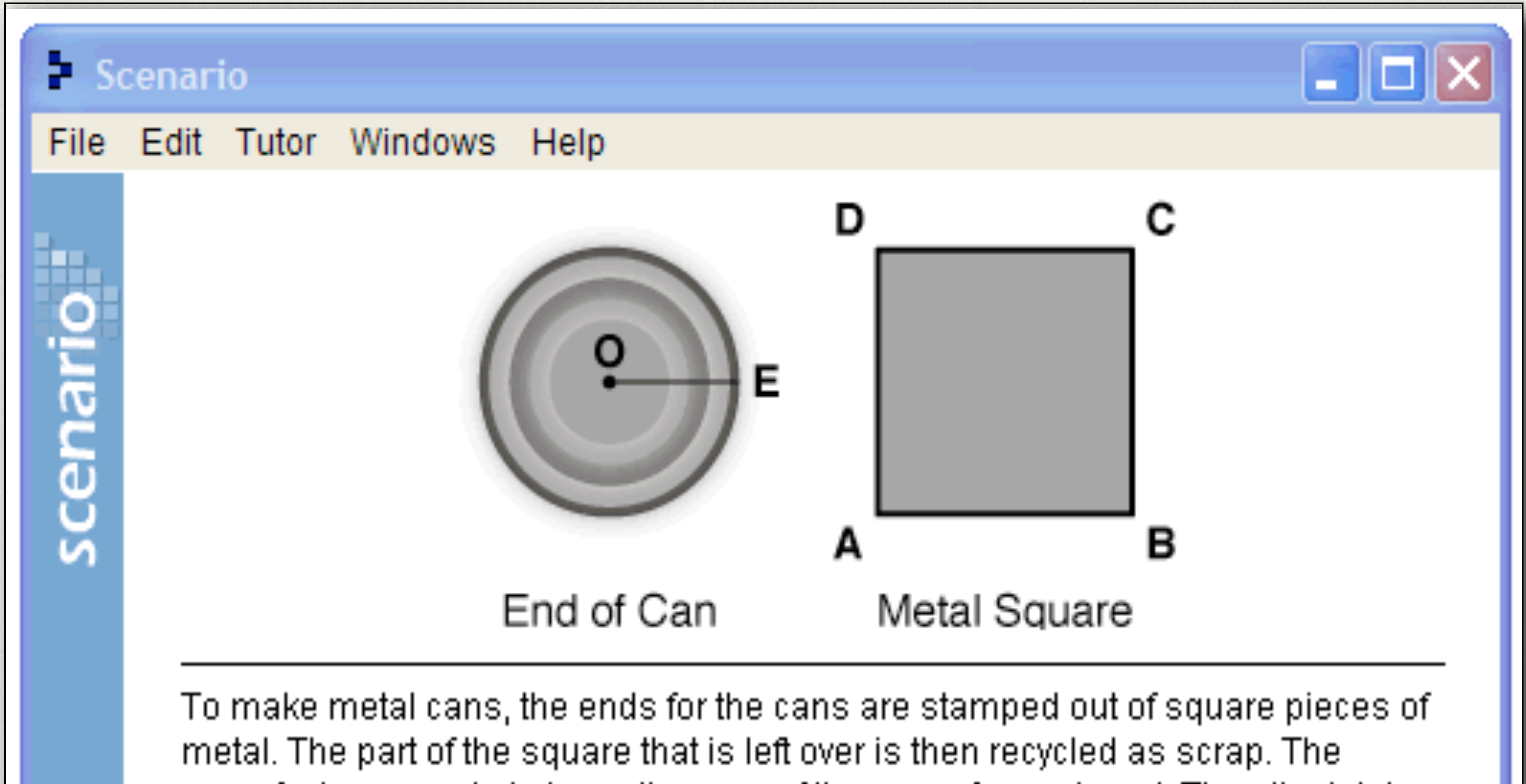
step 17: {A, B}
step 23: {A, C}

[Koedinger, Corbett, Perfetti. Cognitive Science, 2012]

CONSEQUENCES OF KC HYPOTHESIS

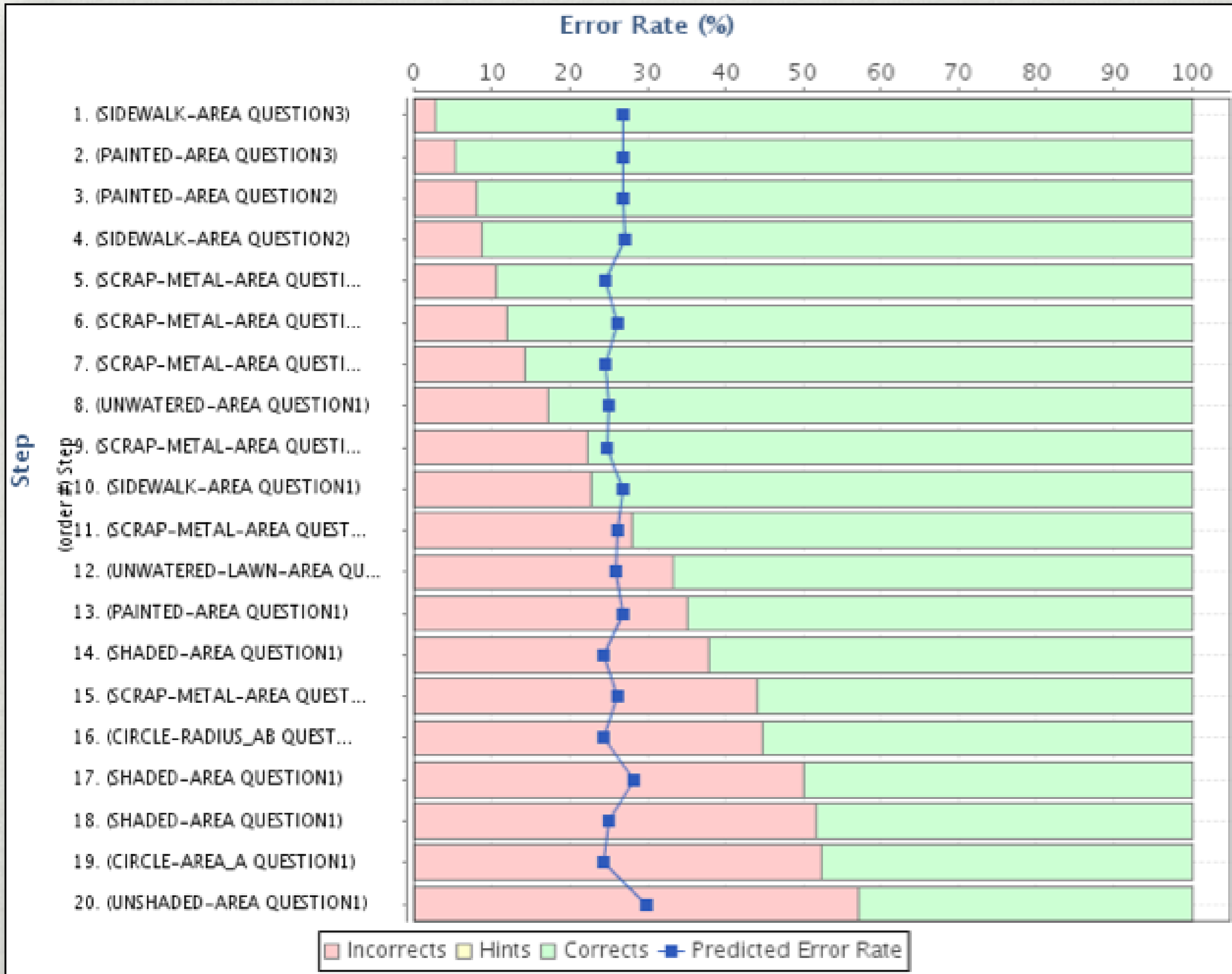
- Mistakes are at KC level: select wrong KC; apply right KC to wrong data; mistake in application of KC
  - identifying the KC at fault makes it easier to give student feedback
- If we can accurately
  - determine list of KCs
  - label instructional activities by KCs
- …then we immediately know the quality/coverage of our content

Compose-by-addition



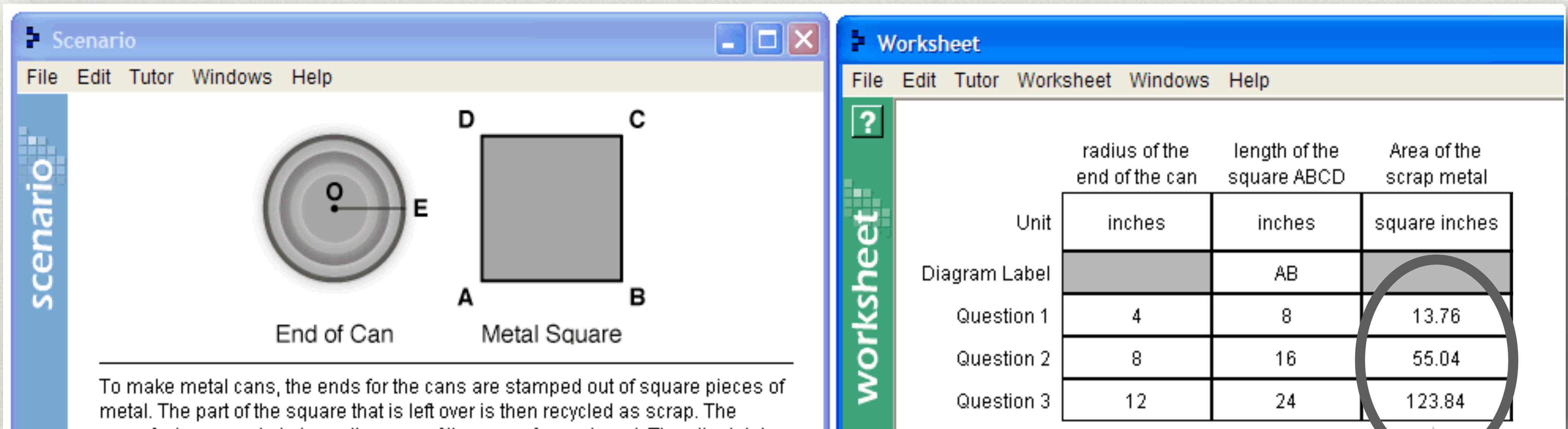
[Stamper & Koedinger, AIED 2011]

COMPOSE-BY-ADDITION

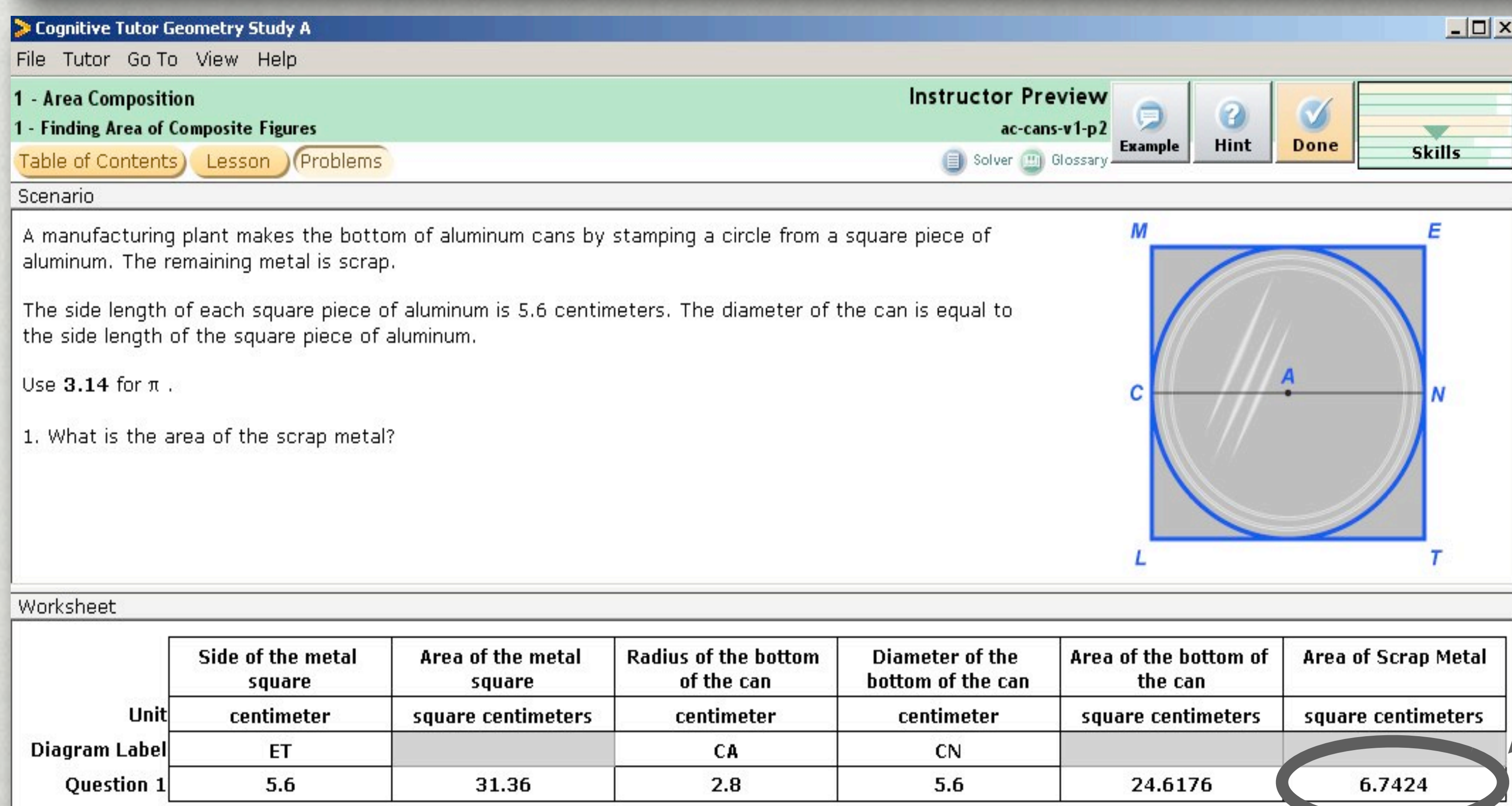


[Stamper & Koedinger, AIED 2011]

WHY ARE SOME COMPOSE-BY-ADDITION STEPS HARDER?



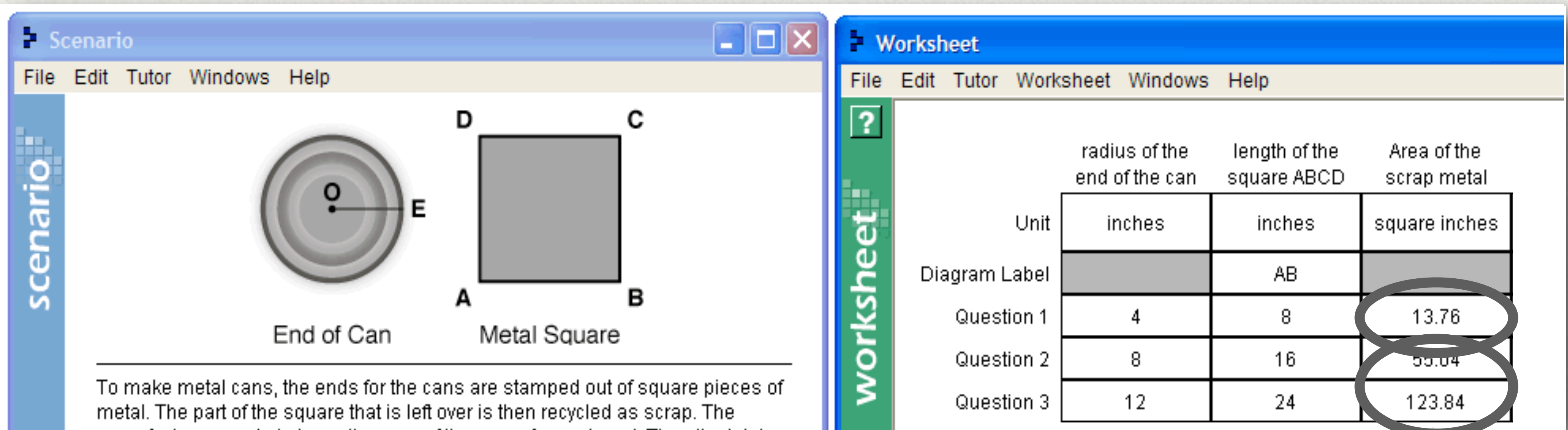
	radius of the end of the can	length of the square ABCD	Area of the scrap metal
Unit	inches	inches	square inches
Diagram Label		AB	
Question 1	4	8	13.76
Question 2	8	16	55.04
Question 3	12	24	123.84



	Side of the metal square	Area of the metal square	Radius of the bottom of the can	Diameter of the bottom of the can	Area of the bottom of the can	Area of Scrap Metal
Unit	centimeter	square centimeters	centimeter	centimeter	square centimeters	square centimeters
Diagram Label	ET		CA	CN		
Question 1	5.6	31.36	2.8	5.6	24.6176	6.7424

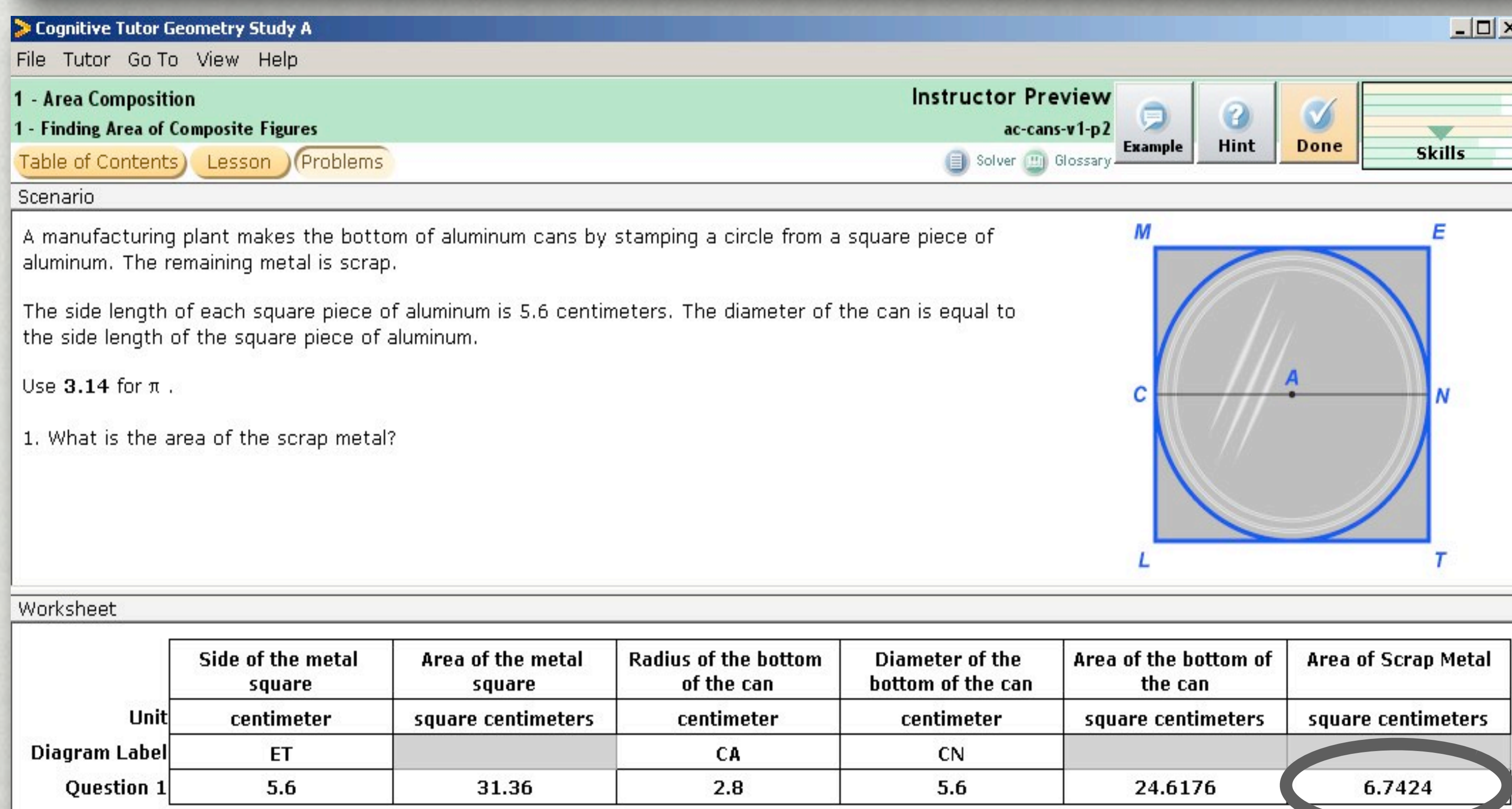
compose by addition

Why are some compose-by-addition steps harder?



hard

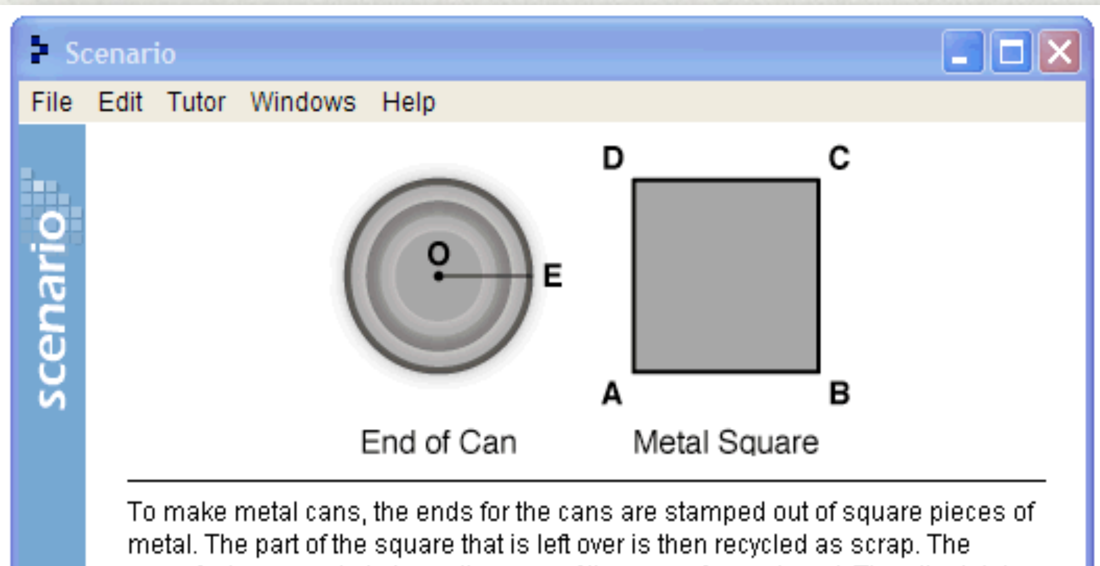
easy



compose by addition

medium

Why are some compose-by-addition steps harder?



Worksheet

File Edit Tutor Worksheet Windows Help

	radius of the end of the can	length of the square ABCD	Area of the scrap metal
Unit	inches	inches	square inches
Diagram Label		AB	
Question 1	4	8	13.76
Question 2	8	16	55.04
Question 3	12	24	123.84

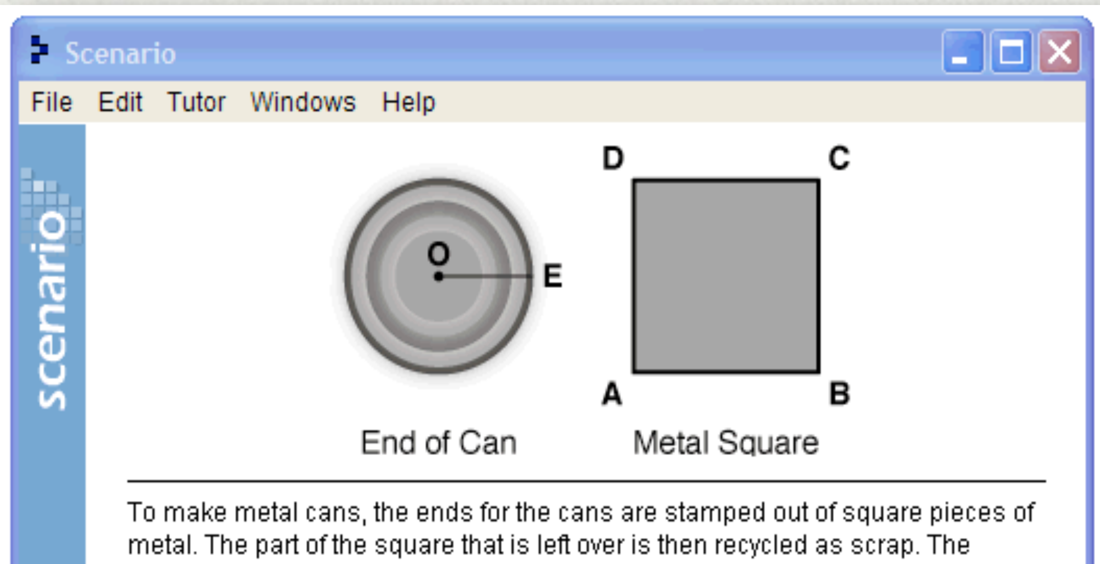
hard

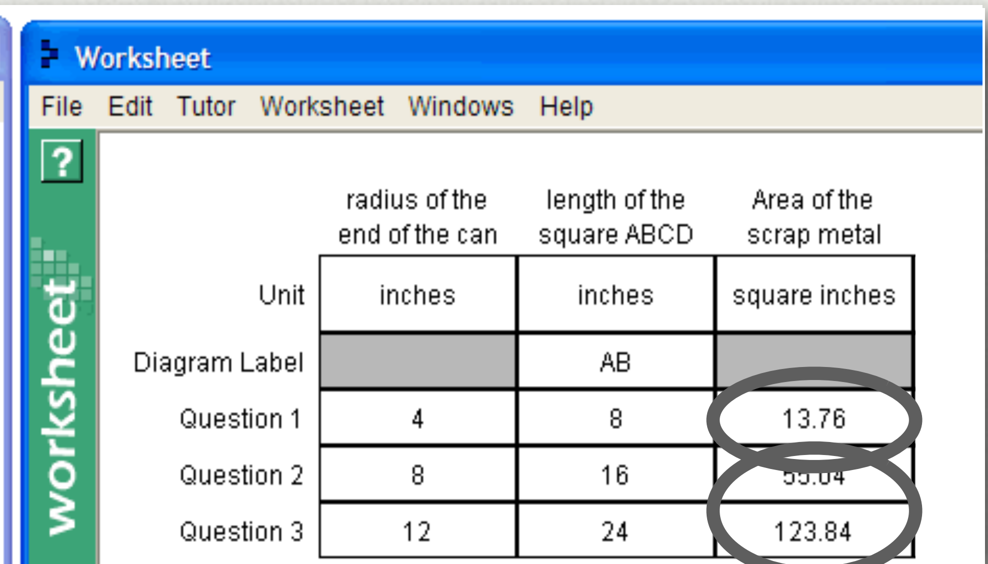
easy

	Side of the metal square	Area of the metal square	Radius of the bottom of the can	Diameter of the bottom of the can	Area of the bottom of the can	Area of Scrap Metal
Unit	centimeter	square centimeters	centimeter	centimeter	square centimeters	square centimeters
Diagram Label	ET		CA	CN		
Question 1	5.6	31.36	2.8	5.6	24.6176	6.7424

medium

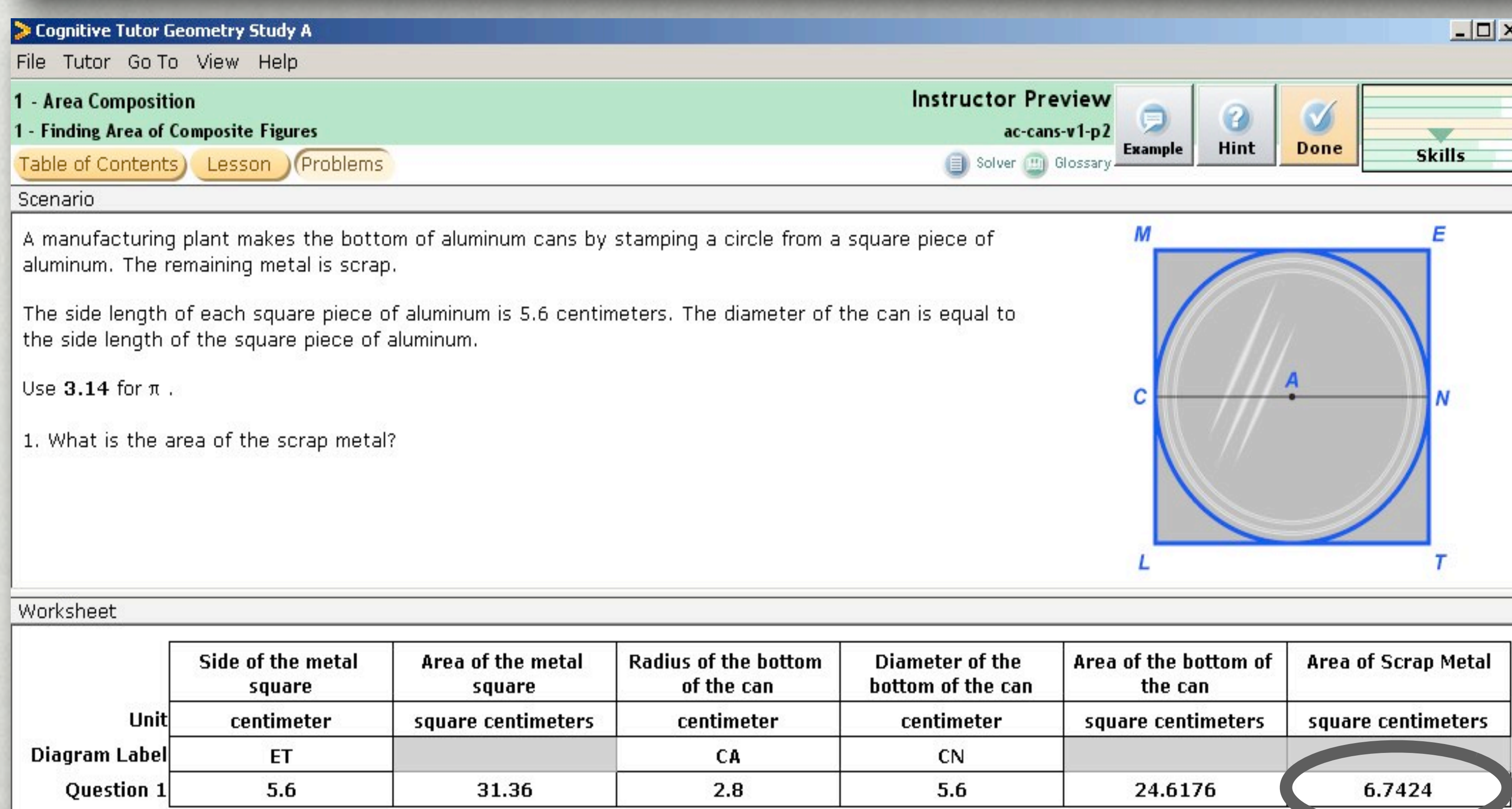
HYPOTHESIS: DIFFERENCE IS IN HOW MUCH PLANNING IS NEEDED





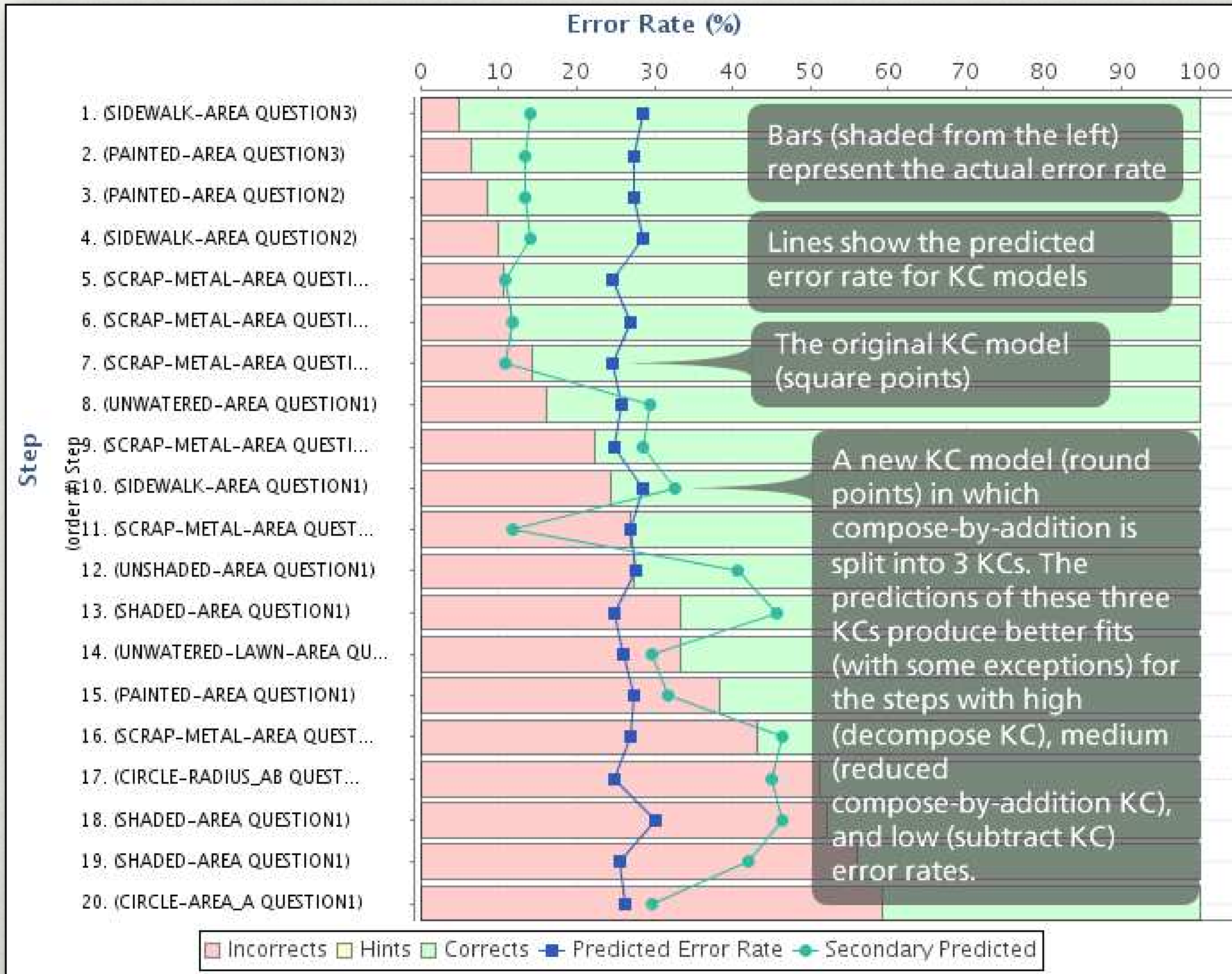
plan to compose

subtract



compose by addition

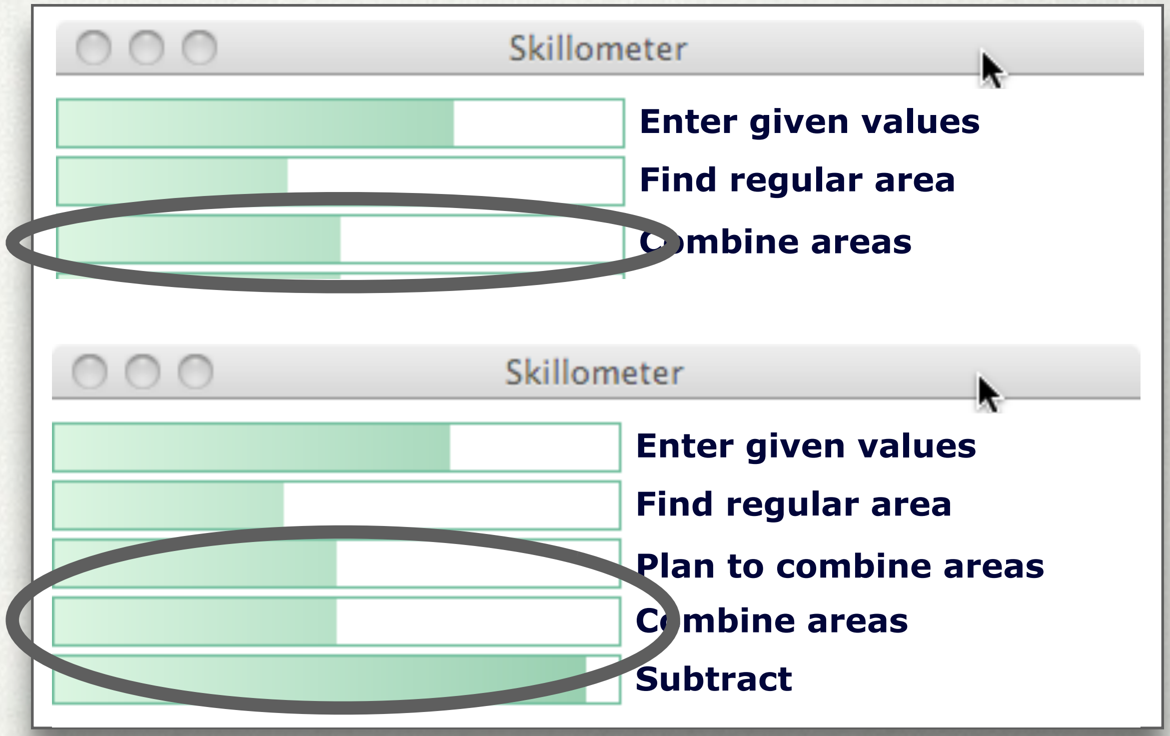
KC DISCOVERY



[Stamper & Koedinger, AIED 2011]

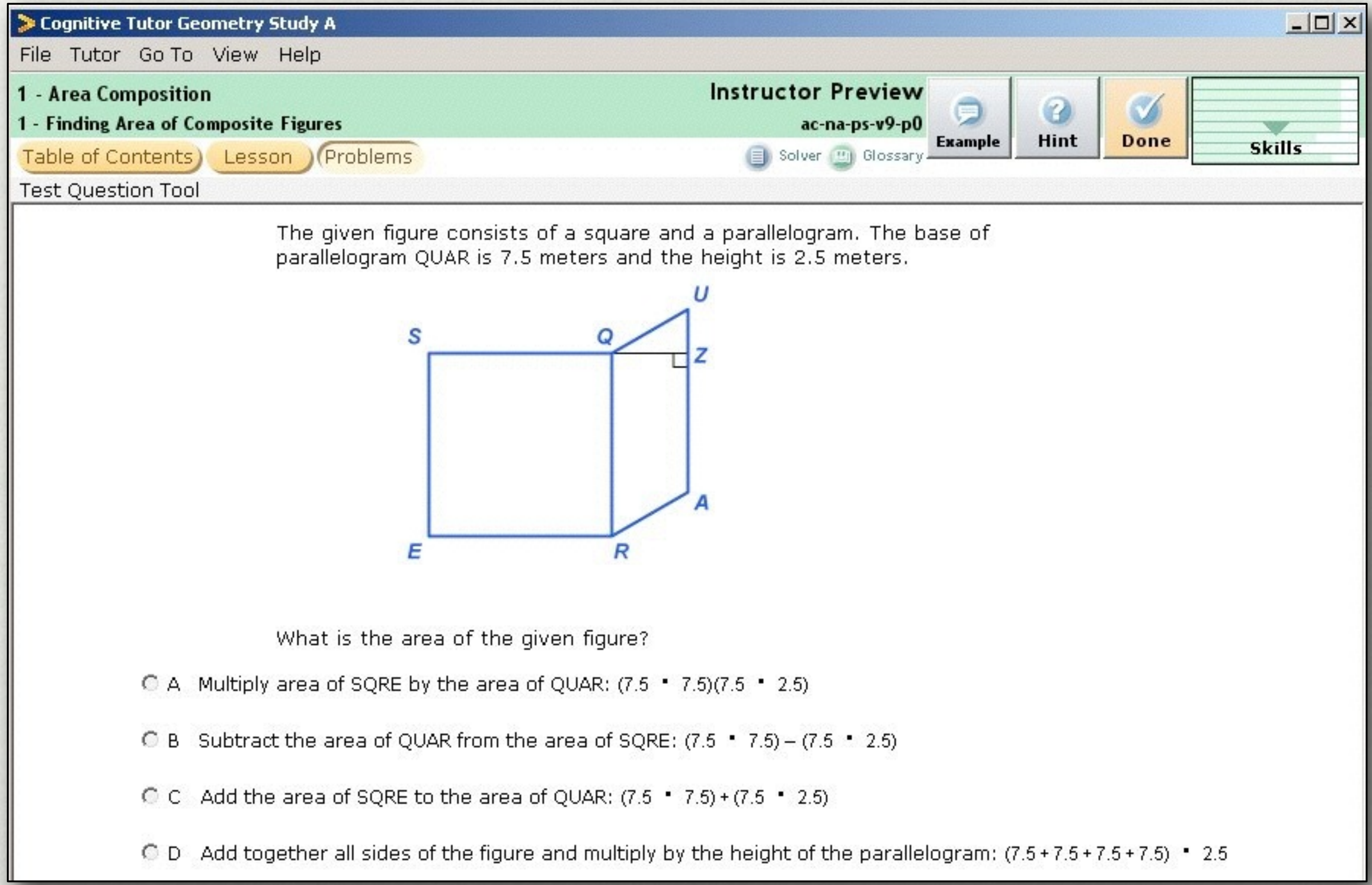
Use data-driven model to redesign tutor

- New skill bars for planning skills
  - skill bars are a tutor interface to show students where they are in acquiring skills
- Sequence for gentle slope
  - adaptive fading of scaffolding
- New problems that focus on planning
  - next slide…



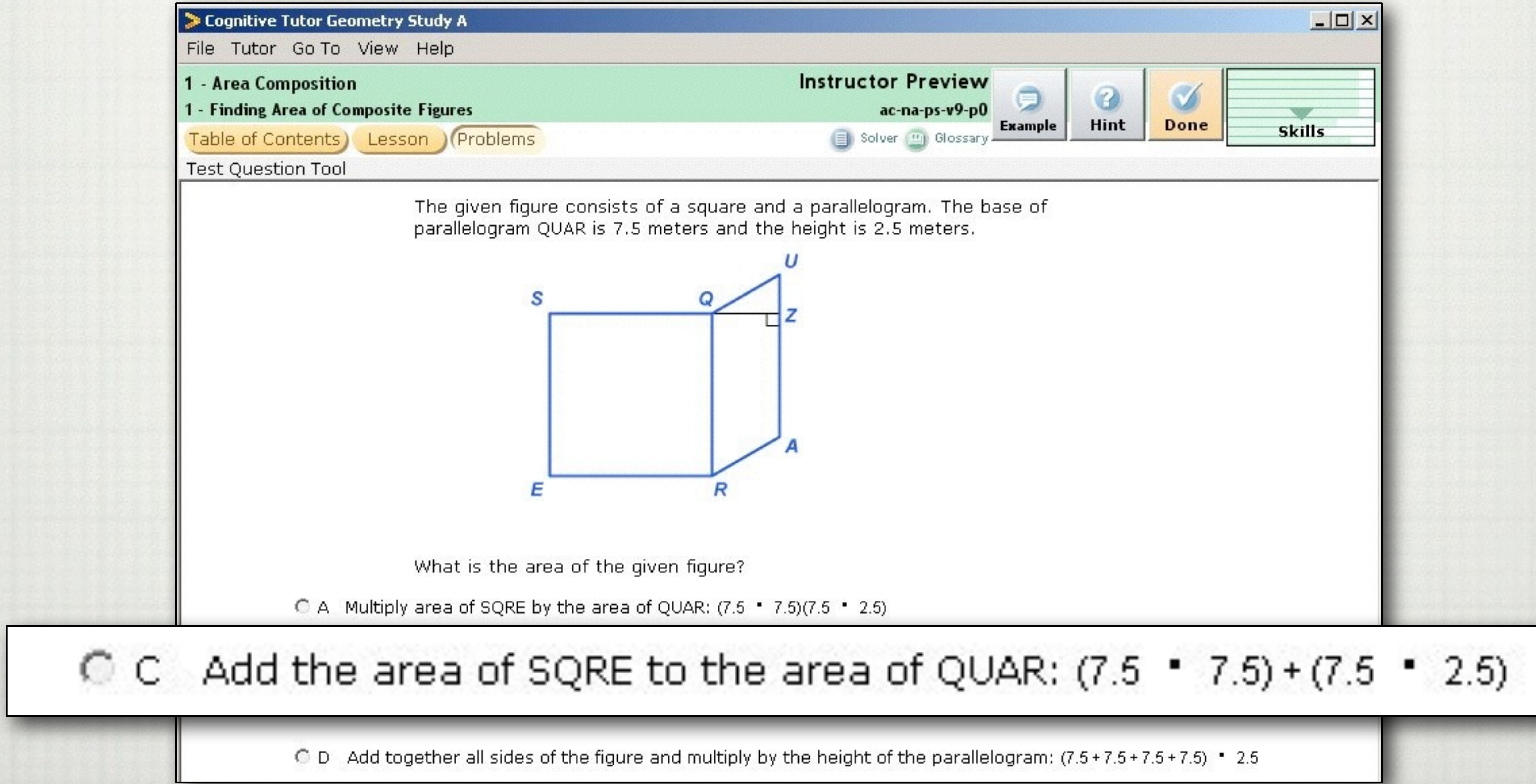
New problems: isolate practice on planning step

- Decompose complex problem into simpler ones



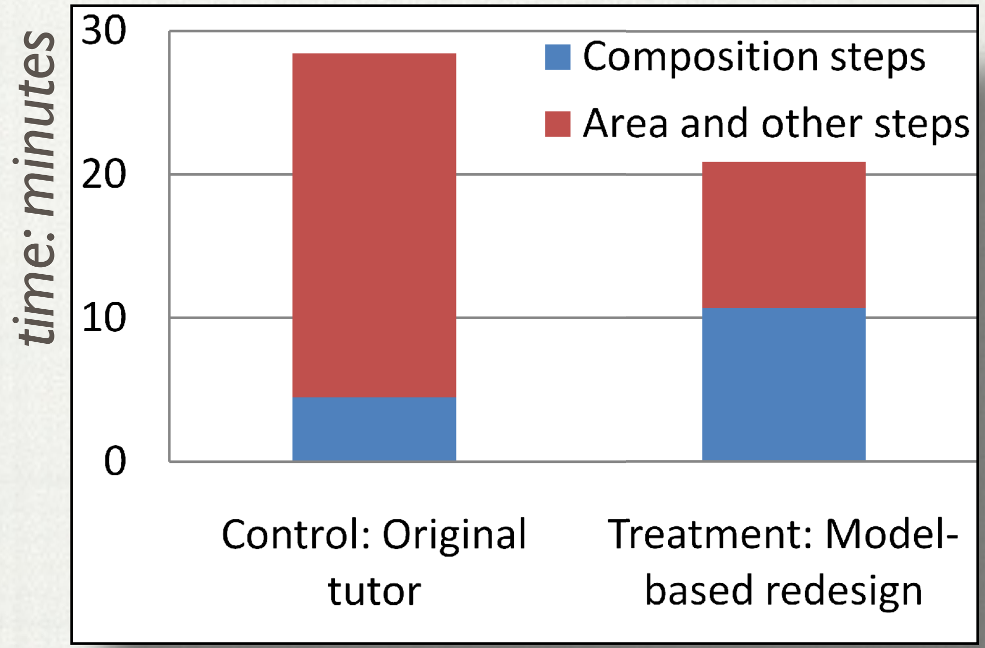
NEW PROBLEMS: ISOLATE PRACTICE ON PLANNING STEP

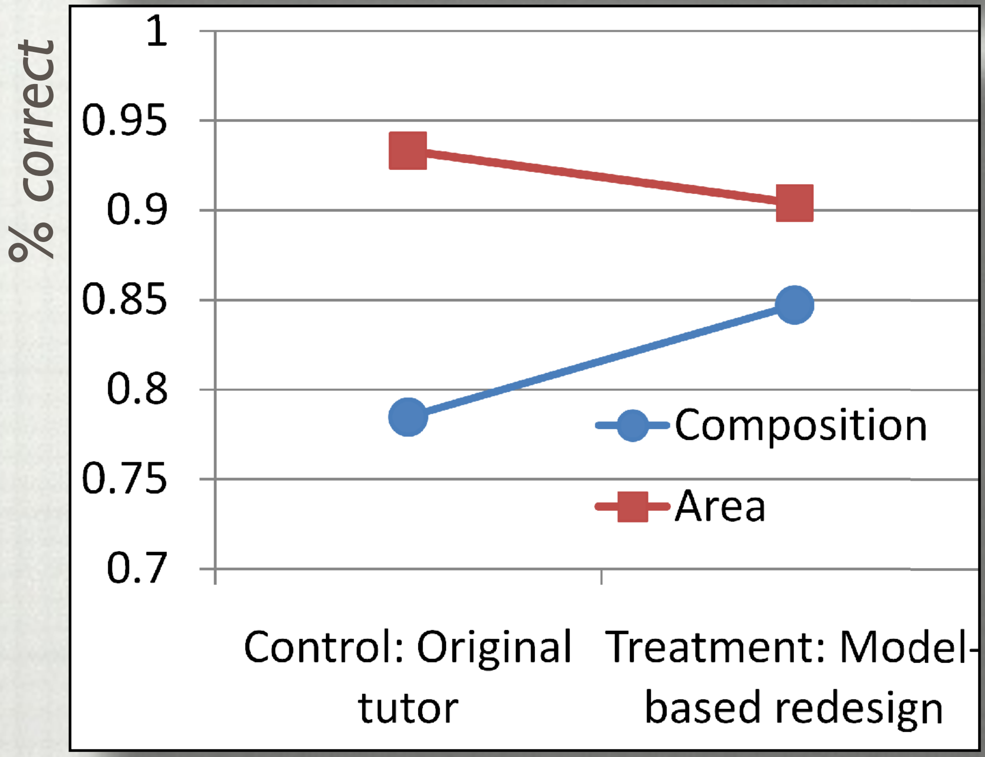
- Decompose complex problem into simpler ones



RESULTS

- More efficient: 25% less student time
  - instructional time by step type
- Better learning of planning skills
  - post-test %correct by item type



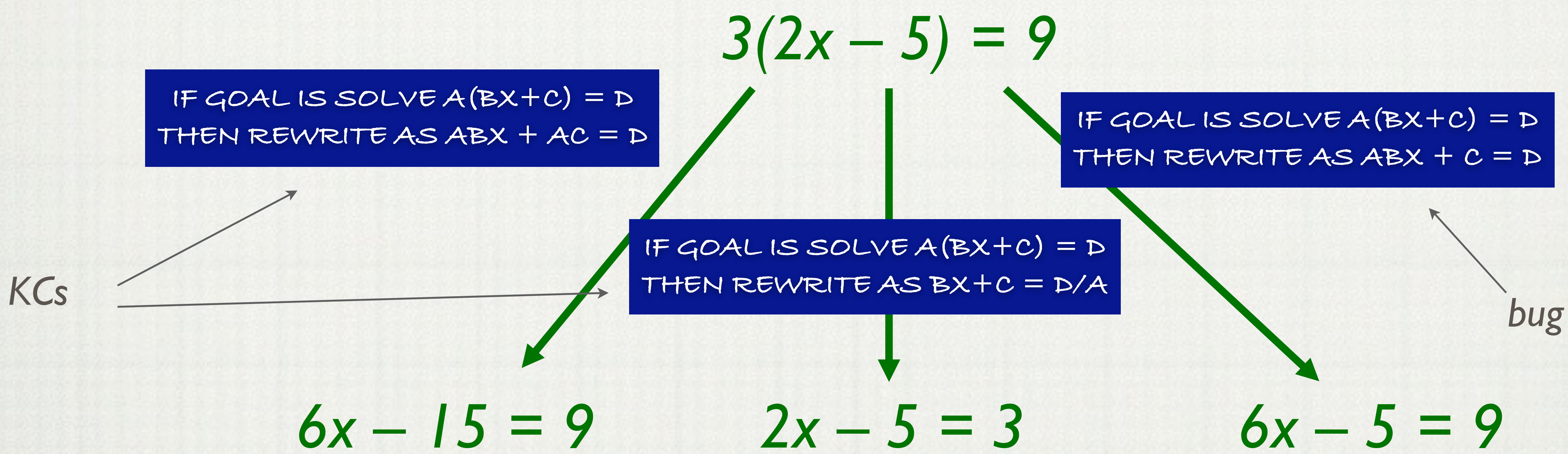


[Stamper & Koedinger, AIED 2011]

More structure: what's in a KC?

- So far, each KC is just present or absent in a student or problem
- Nothing to distinguish algebra KCs from ESL KCs
- What's going on under the hood as a student solves a problem?

Rule-based Cognitive Model



- What does it look like inside the student's brain?
  - … maybe a rule-based system
  - … in which case KCs might correspond to rules
    - :- president of US is Obama
    - constant C on LHS of equation E :- move C to RHS of E

Rule-based system

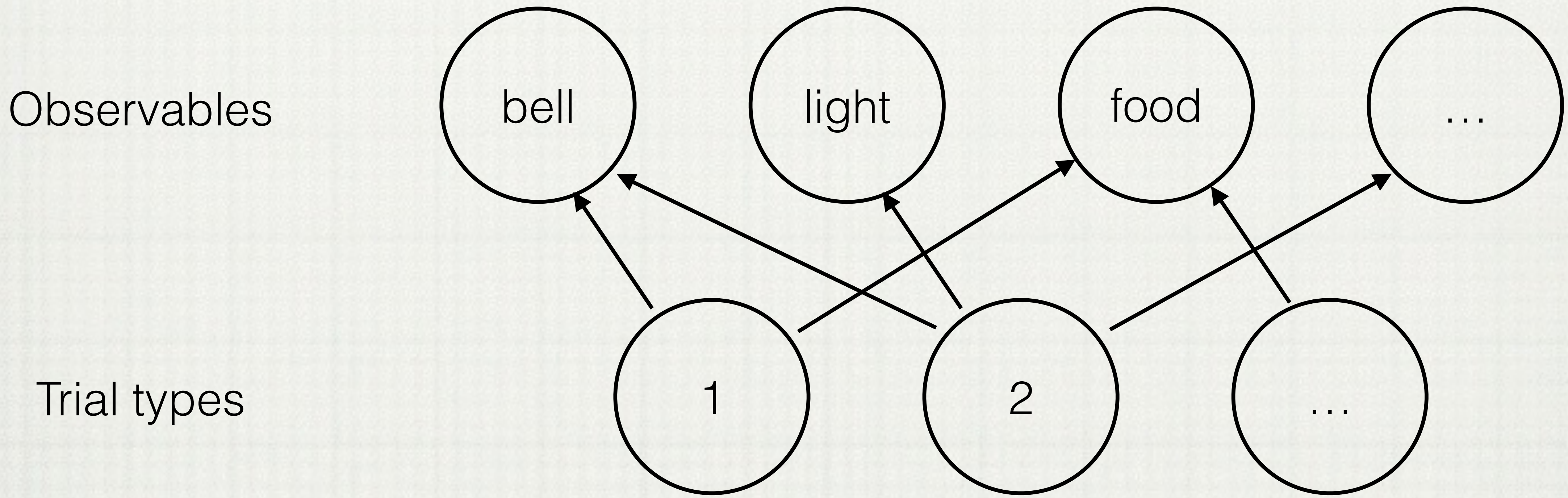
- Aka production system:
  - declarative knowledge held in working memory — "I see 3x+5 = 8"
  - production rules match declarative knowledge — "if LHS has constant **C**…"
  - and act on WM or external world — "… then subtract **C** from both sides"
- Much cognitive modeling work endorses this claim explicitly or implicitly
  - ACT-R, SimStudent, Russell & Norvig, …
- But two problems: uncertainty handling, representation learning
  - here's where more ML research can help!

Problem 1: uncertainty

A day in the life of a rat

Trial	Bell?	Light?	Food?
1	×	✓	✓
2	✓	×	×
3	×	✓	×
4	×	✓	✓
…	…	…	…

Rat as Bayesian



- Priors over: how many trial types, sparsity of connections, reliability of connections, …
- (This is a common architecture for medical diagnosis systems)

Quiz: are you smarter than a rat?

Trial	Bell?	Light?	Food?
1	×	✓	✓
2	×	✓	×
3	×	✓	✓
4	×	✓	✓
…	…	…	…
100	×	✓	✓

Quiz: are you smarter than a rat?

Trial	Bell?	Light?	Food?
1	×	✓	✓
2	×	✓	×
3	×	✓	✓
4	×	✓	✓
…	…	…	…
100	×	✓	✓
101	✓	✓	×

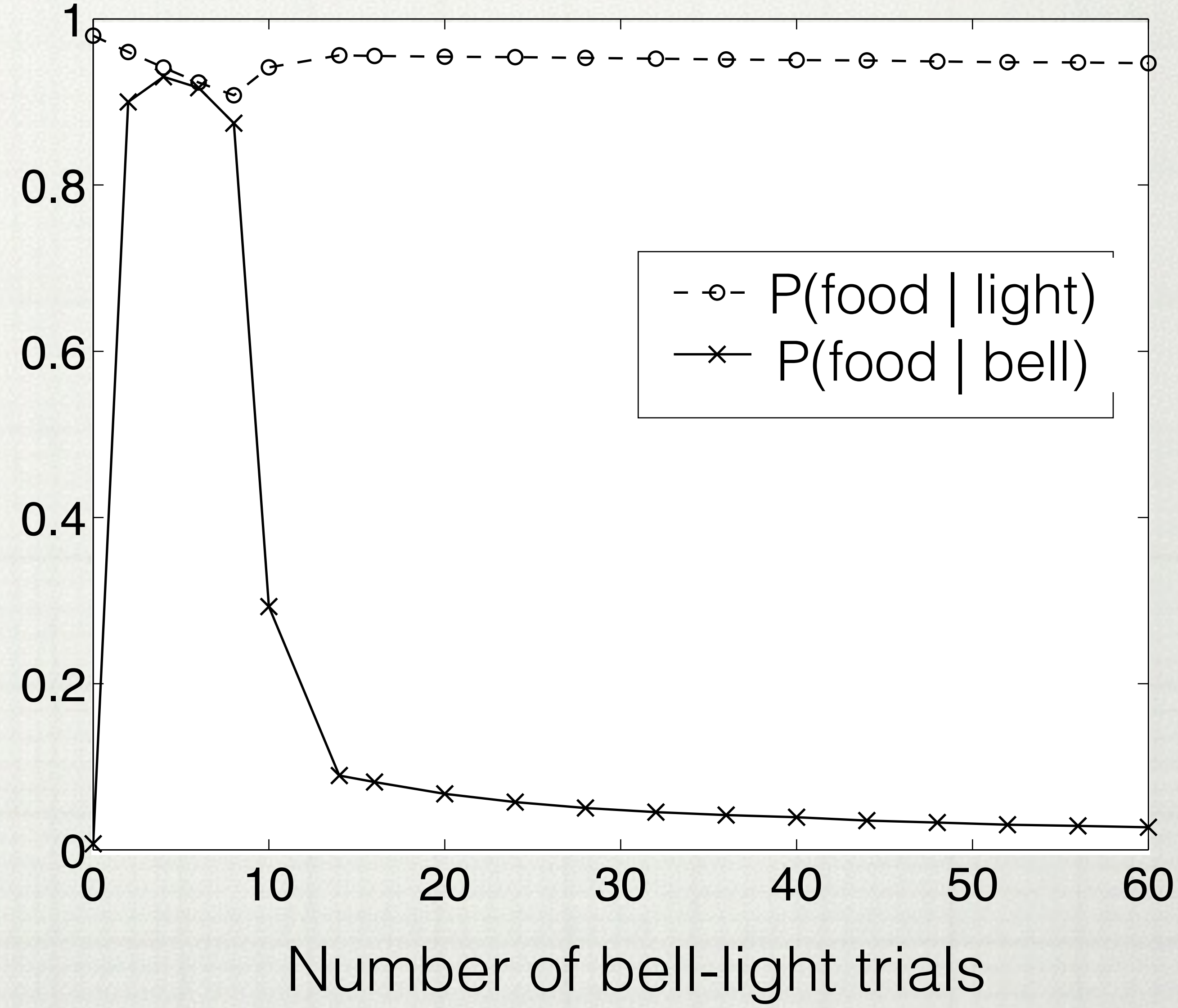
AND THE RAT SAYS...

Effect name	2nd-order conditioning	Conditioned inhibition
light-food trials	many	many
bell-light trials	few	many
test: bell predicts food?	↑	↓

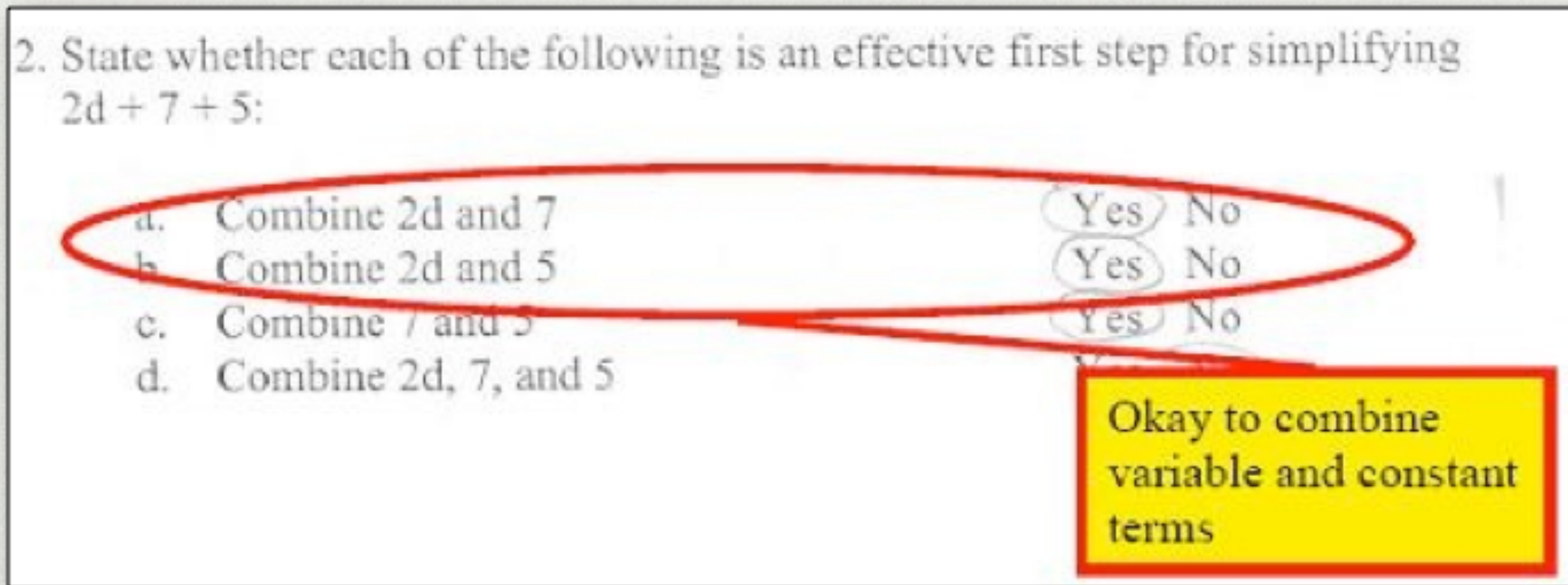
- Both right! With more light-bell trials, evidence increases for a separate trial type.

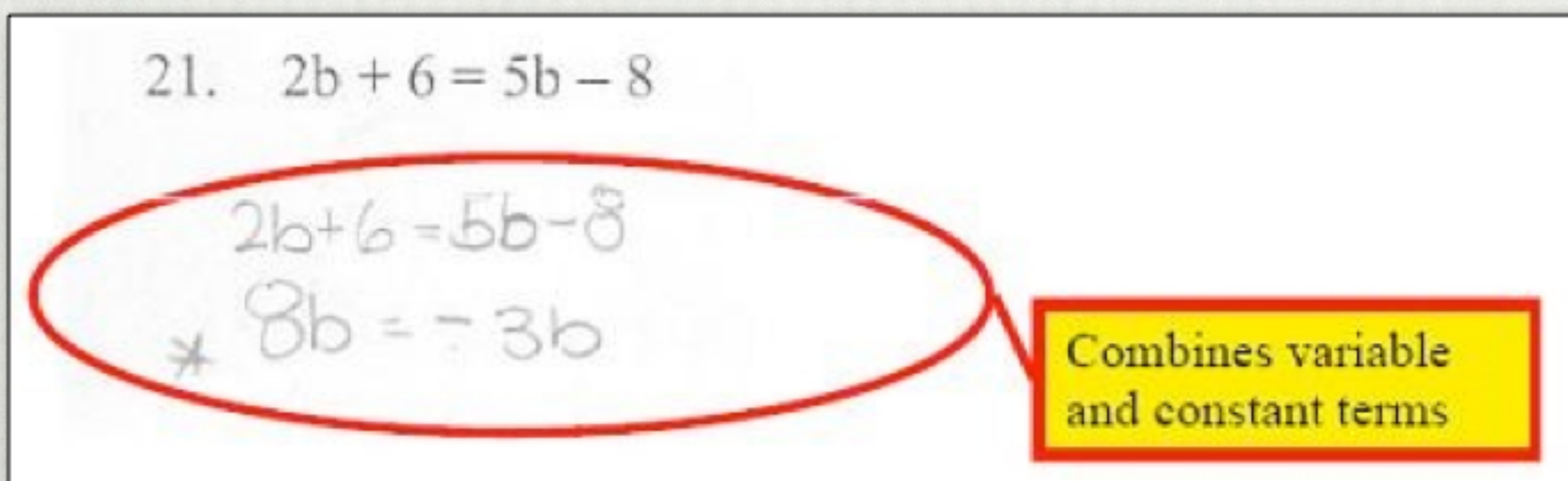
Bayesian rule learning in classical conditioning

- Only fully Bayesian inference/learning captured both effects *[Courville, Daw, Gordon, Touretzky, NIPS 2003]*
- Few bell-light trials, 1 trial type: (bell, light, food) all associated
- More trials: (bell, light, no food) v. (light, food, no bell)



PROBLEM 2: REPRESENTATION LEARNING





Flaw with "KC = rule": Many bugs come from weak features

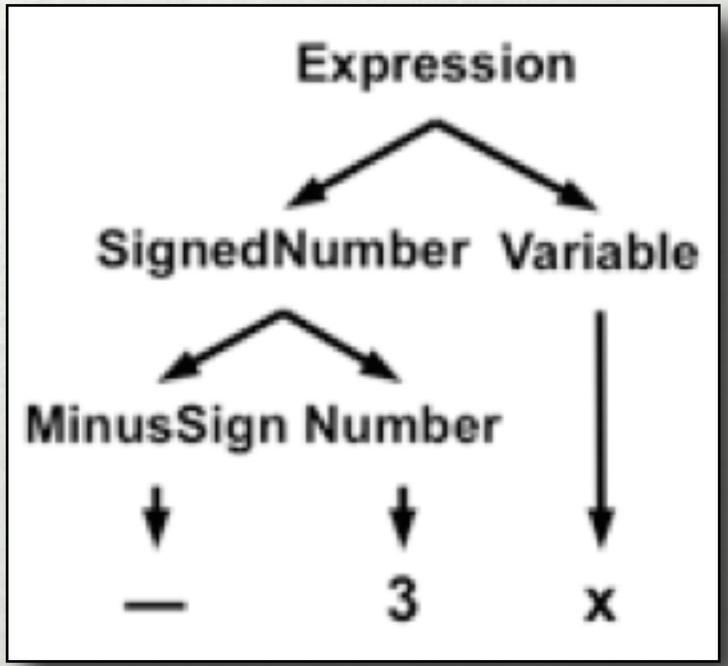
Representation learning

- Some student errors come from failure to correctly interpret (internally represent) a problem
- As student sees more and more examples like $3x + 5 = 8$, gets better and better “language model” to explain them (build internal representation)
- —> some KCs must correspond to features of the improved language model

Experiment

- Present algebra examples to a machine learning system
- As part of learning, induce a language model (an unsupervised probabilistic context free grammar) for algebra equations
- Make output of language model (grammar nonterminals, e.g., SignedNumber) available as features of each example
- Use these features in simulated problem-solving to discover KCs

[Li, Cohen, Koedinger, Matsuda, 2010]



NEW COGNITIVE MODELS ARE MORE ACCURATE

Table 8.3: CV RMSE on SimStudent-Generated models and Human-Generated Models.

	Human-Generated Model	SimStudent-Discovered Model
Algebra	0.4024	**0.3999**
Stoichiometry	0.3501	**0.3488**
Fraction Addition	**0.3232**	0.3343
Article Selection	0.4044	**0.4033**

[Li, Cohen, Koedinger, Matsuda, 2010]

Open research question

- Can we build a new generation of rule-based system that has
  - rich uncertainty handling
  - integrated representation learning
- … and use it to help us model student learning?

SUMMARY

- A key contribution of machine learning to education will be to help understand the educational content we're creating and delivering
- Essential idea: ML models of structured latent variables
- Specifically, build and test hypotheses about the **knowledge**, **procedures** and **representations** students use to **solve problems**
  - latents = KCs, rules, representations, strategies, …
- Need to link uncertainty handling (traditional domain of ML) to new, harder situations encountered in understanding student knowledge
- Exciting time for research in ML and education!