Learning from Learning Curves: Item Response Theory & Learning Factors Analysis

Ken Koedinger
Human-Computer Interaction Institute
Carnegie Mellon University

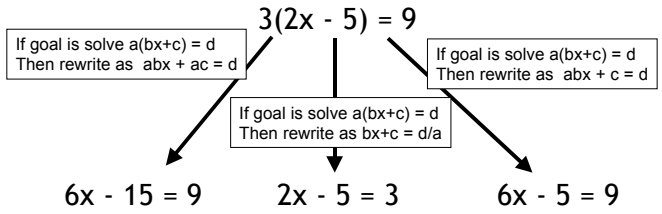
Cen, H., Koedinger, K., Junker, B. **Learning Factors Analysis - A General Method for Cognitive Model Evaluation and Improvement**. *8th International Conference on Intelligent Tutoring Systems*. 2006.

Stamper, J. & Koedinger, K.R. **Human-machine student model discovery and improvement using data**. *Proceedings of the 15th International Conference on Artificial Intelligence in Education*. 2011.

Cognitive Tutor Technology

Use cognitive model to individualize instruction

- Cognitive Model: A system that can solve problems in the various ways students can

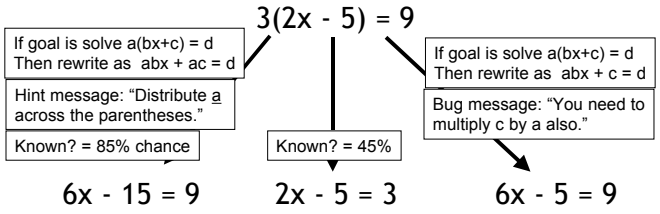


- Model Tracing: Follows student through their individual approach to a problem -> context-sensitive instruction

Cognitive Tutor Technology

Use cognitive model to individualize instruction

- Cognitive Model: A system that can solve problems in the various ways students can



- Model Tracing: Follows student through their individual approach to a problem -> context-sensitive instruction
- Knowledge Tracing: Assesses student's knowledge growth -> individualized activity selection and pacing

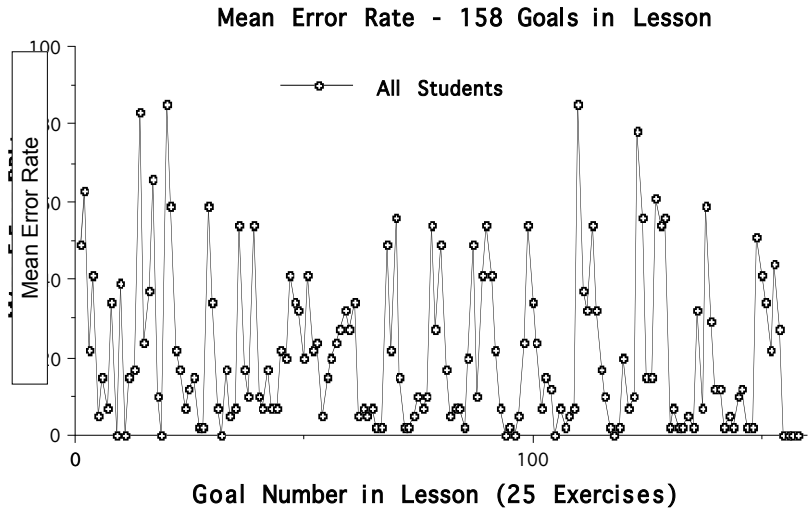
Cognitive Model Discovery

- Traditional Cognitive Task Analysis
  - Interview experts, think alouds, DFA
- Result: cognitive model of student knowledge
  - Cognitive model drives ITS behaviors & instructional design decisions
- Key goal for Educational Data Mining
  - Improve Cognitive Task Analysis
  - Use student data from initial tutor
  - Employ machine learning & statistics to discover better cognitive models

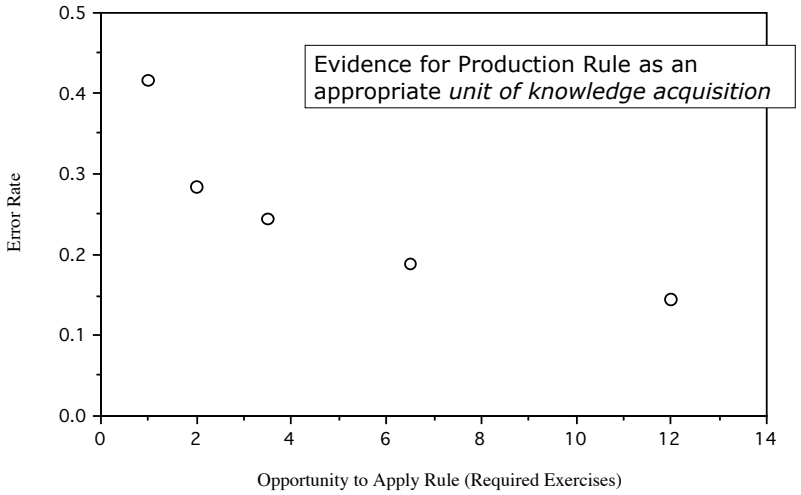
Overview

- Using *learning curves* to evaluate cognitive models
- Statistical models of student performance & learning
  - Example of improving tutor
  - Comparison to other Psychometric models
- Using *Learning Factors Analysis* to discover better cognitive models
- Educational Data Mining research challenges

Student Performance As They Practice with the LISP Tutor



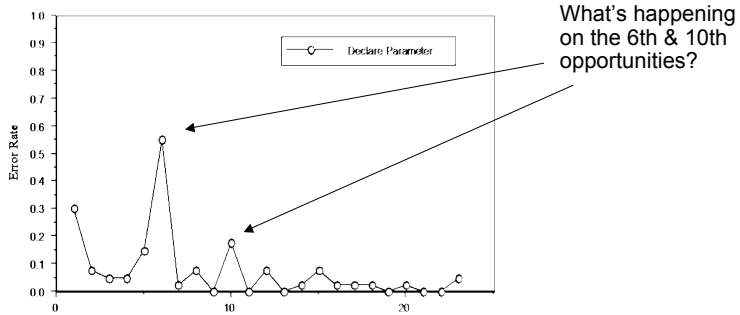
Production Rule Analysis



Using learning curves to evaluate a cognitive model

- Lisp Tutor Model
  - Learning curves used to validate cognitive model
  - Fit better when organized by knowledge components (productions) rather than surface forms (programming language terms)
- But, curves not smooth for some production rules
  - "Blips" in leaning curves indicate the knowledge representation may not be right
    - Corbett, Anderson, O'Brien (1995)
  - Let me illustrate …

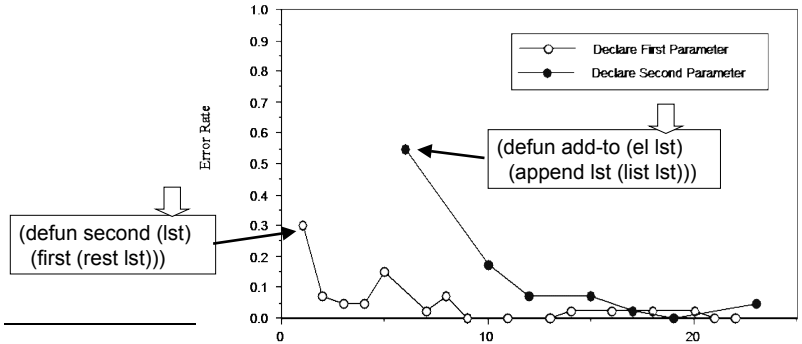
Curve for "Declare Parameter" production rule



- How are steps with blips different from others?
- What's the unique feature or *factor* explaining these blips?

Can modify cognitive model using unique *factor* present at "blips"

- Blips occur when to-be-written program has 2 parameters
- *Split* Declare-Parameter by parameter-number factor:
  - Declare-first-parameter
  - Declare-second-parameter



Can learning curve analysis be automated?

- Manual learning curve analysis
  - Identify "blips" in learning curve visualization
  - Manually create a new model
  - Qualitative judgment of fit
- Toward automatic learning curve analysis
  - Blips as deviations from statistical model
  - Propose alternative cognitive models
  - Evaluate cognitive model using prediction accuracy statistics

Overview

- Using *learning curves* to evaluate cognitive models
- Statistical models of student performance & learning
  - Example of improving tutor
  - Comparison to other Psychometric models
- Using *Learning Factors Analysis* to discover better cognitive models
- Educational Data Mining research challenges

Representing Knowledge Components as factors of items

- Problem: How to represent KC model?
- Solution: Q-Matrix (Tatsuoka, 1983)

Items X Knowledge Components (KCs)

Item | KCs:	Add	Sub	Mul	Div
2*8	0	0	1	0
2*8 - 3	0	1	1	0

- Single KC item = when a row has one 1
- Multi-KC item = when a row has many 1's

Q matrix is a bridge between a symbolic cognitive model & a statistical model

Additive Factors Model Assumptions

- Logistic regression to fit learning curves (Draney, Wilson, Pirolli, 1995)
- Assumptions about knowledge components (KCs) & students
  - Different students may initially know more or less
  - Students generally learn at the same rate
  - Some KCs are initially easier than others
  - Some KCs are easier to learn than others
- These assumptions are reflected in a statistical model
  - Intercept parameters for each student
  - Intercept & slope parameters for each KC
    - Slope = for every practice opportunity there is an increase in predicted performance

Simple Statistical Model of Performance & Learning

- Problem: How to predict student responses from model?
- Solution: Additive Factor Model
  - i students, j problems/items, k knowledge components (KCs)

Correct?

Step j uses skill k

$$\log \frac{p_{ij}}{1-p_{ij}} = \theta_i + \sum_k \beta_k Q_{kj} + \sum_k Q_{kj}(\gamma_k T_{ik})$$

Opp count

Model parameters:

Student intercept

KC intercept

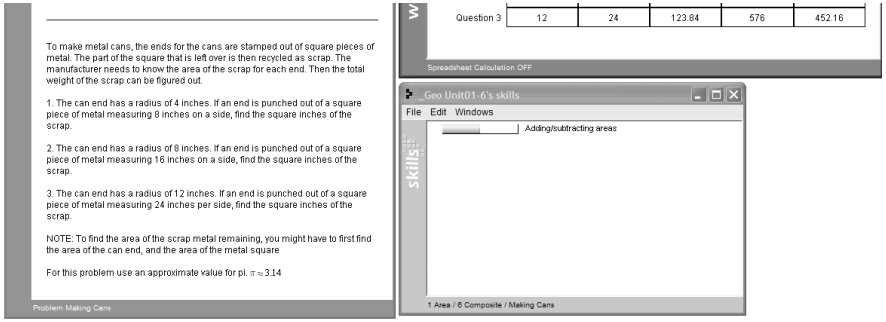
KC slope

Area Unit of Geometry Cognitive Tutor

- Original cognitive model in tutor:

15 skills:

Circle-area
Circle-circumference
Circle-diameter
Circle-radius
Compose-by-addition
Compose-by-multiplication
Parallelogram-area
Parallelogram-side
Pentagon-area
Pentagon-side
Trapezoid-area
Trapezoid-base
Trapezoid-height
Triangle-area
Triangle-side



Log Data Input to AFM

Items = steps in tutors with step-based feedback

Q-matrix in single column: works for single KC items

Opportunities Student has had to learn KC

Student	Step (Item)	KC	Opportunity	Success
A	p1s1	Circle-area	0	0
A	p2s1	Circle-area	1	1
A	p2s2	Rectangle-area	0	1
A	p2s3	Compose-by-addition	0	0
A	p3s1	Circle-area	2	0

AFM Results for original KC model

Higher intercept of skill -> easier skill

Higher slope of skill -> faster students learn it

Skill	Intercept	Slope	Avg Opportunties	Initial Probability	Avg Probability	Final Probability
Parallelogram-area	2.14	-0.01	14.9	0.95	0.94	0.93
Pentagon-area	-2.16	0.45	4.3	0.2	0.63	0.84

Student	Intercept
student0	1.18
student1	0.82
student2	0.21

Higher intercept of student -> student initially knew more

Model Statistics	
AIC	3,950
BIC	4,285
MAD	0.083

The AIC, BIC & MAD statistics provide alternative ways to evaluate models

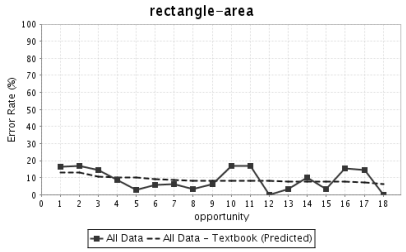
MAD = Mean Absolute Deviation

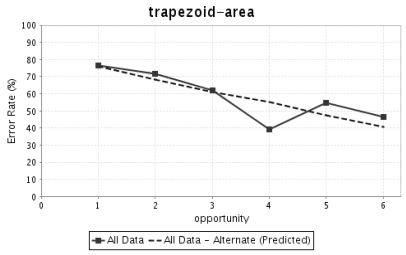
Overview

- Using *learning curves* to evaluate cognitive models
- Statistical models of student performance & learning
  - Example of improving tutor
  - Comparison to other Psychometric models
- Using *Learning Factors Analysis* to discover better cognitive models
- Educational Data Mining research challenges

Application: Use Statistical Model to improve tutor

- Some KCs over-practiced, others under (Cen, Koedinger, Junker, 2007)



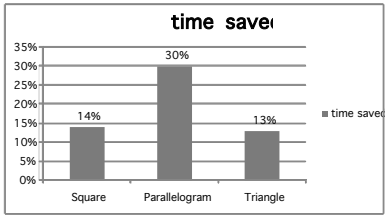


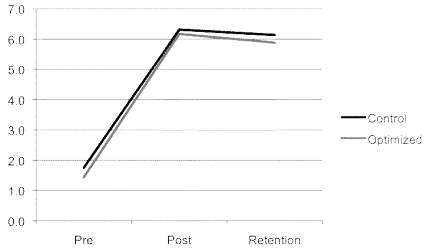
initial error rate 12%
reduced to 8%
after 18 times of practice

initial error rate 76%
reduced to 40%
after 6 times of practice

"Close the loop" experiment

- *In vivo* experiment: New version of tutor with updated knowledge tracing parameters vs. prior version
- Reduced learning time by 20%, same robust learning gains
- Knowledge transfer: Carnegie Learning using approach for other tutor units





Additive Factor Model (AFM) generalizes Item Response Theory (IRT)

- Instance of logistic regression
  - Example: In R use generalized linear regression with family=binomial
  - glm(prob-correct ~ student + KC + KC:opportunity, family=binomial,…)
- Generalization of item response theory (IRT)
  - IRT simply has i student & j item parameters
  - glm(prob-correct ~ student + item, family=binomial,…)
- AFM is different from IRT because:
  - It clusters items by knowledge components
  - It has an opportunity *slope* for each KC

Correct?

Step j uses skill k

$$\log \frac{p_{ij}}{1-p_{ij}} = \theta_i + \sum_k \beta_k Q_{kj} + \sum_k Q_{kj}(\gamma_k T_{ik})$$

Opp count

Comparing to other psychometric models

- AFM adds a growth component to "LLTM" (Wilson & De Boeck)
  - LTTM is an "item explanatory" generalization of IRT or "Rasch"
  - "Person explanatory" models are related to factor analysis and

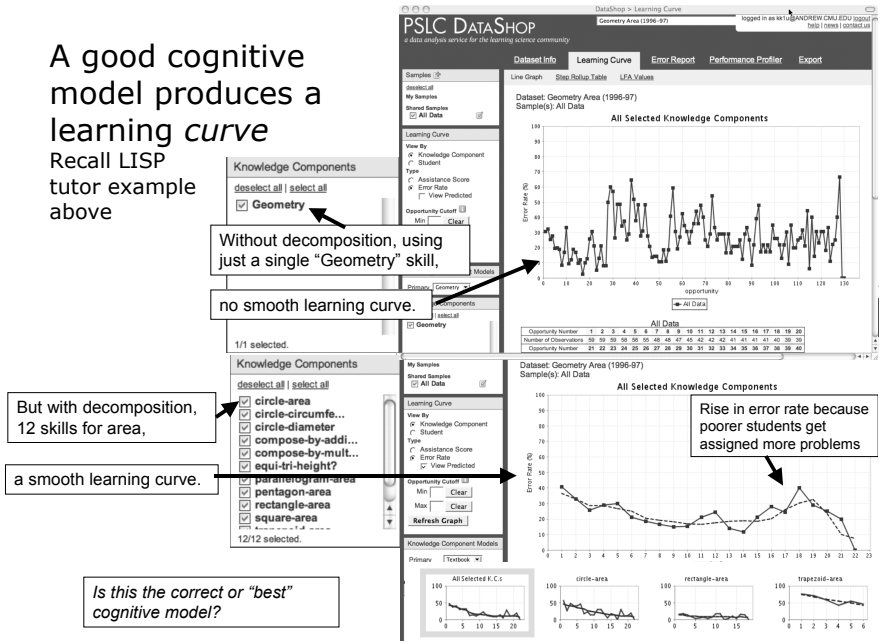
	$\eta_{pi} =$			
Model	Person part	Item part	Random effect	Model type
Rasch model	θ_p	$-\beta_i$	$\theta_p \sim N(0, \sigma_\theta^2)$	Doubly descriptive
Latent reg Rasch model	$\sum_{j=1}^{J} \vartheta_j Z_{pj} + \varepsilon_p$	$-\beta_i$	$\varepsilon_p \sim N(0, \sigma_\varepsilon^2)$	Person explanatory
LLTM	θ_p	$-\sum_{k=0}^{K} \beta_k X_{ik}$	$\theta_p \sim N(0, \sigma_\theta^2)$	Item explanatory
Latent reg LLTM	$\sum_{j=1}^{J} \vartheta_j Z_{pj} + \varepsilon_p$	$-\sum_{k=0}^{K} \beta_k X_{ik}$	$\varepsilon_p \sim N(0, \sigma_\varepsilon^2)$	Doubly explanatory

Model Evaluation

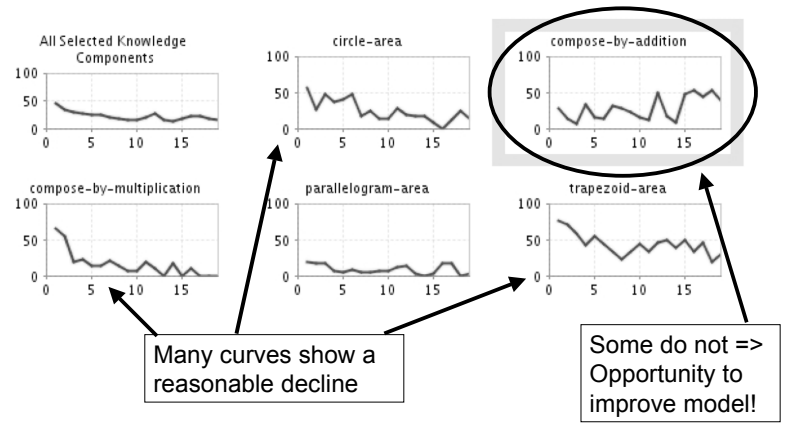
- How to compare cognitive models?
  - A good model minimizes prediction risk by balancing fit with data & complexity (Wasserman 2005)
- Model-data *fit* metrics
  - Log likelihood, root mean squared error (RMSE), mean average deviation (MAD), area under curve (AUC), …
- *Prediction* metrics
  - BIC, AIC: Faster metrics add a penalty for # parameters
    - BIC = -2*log-likelihood + numPar * log(numOb)
  - Cross validation: Slower but better
    - Split data in training & test sets, optimize parameters with training set, apply fit metrics on test set

A good cognitive model produces a learning *curve*

Recall LISP tutor example above



DataShop visualizations to aid "blip" detection



Learning Factors Analysis

Overview

- Using *learning curves* to evaluate cognitive models
- Statistical models of student performance & learning
  - Example of improving tutor
  - Comparison to other Psychometric models
- Using *Learning Factors Analysis* to discover better cognitive models
- Educational Data Mining research challenges

Learning Factors Analysis (LFA): A Tool for Cognitive Model Discovery

- LFA is a method for discovering & evaluating alternative cognitive models
  - Finds knowledge components that best predict student performance & learning transfer
- Inputs
  - Data: Student success on tasks in domain over time
  - Codes: Factors hypothesized to drive task difficulty & transfer
- Outputs
  - A rank ordering of most predictive cognitive models
  - Parameter estimates for each model

Learning Factors Analysis (LFA) draws from multiple disciplines

- Cognitive Psychology
  - Learning curve analysis (Corbett, et al 1995)
- Psychometrics & Statistics
  - Q Matrix & Rule Space (Tatsuoka 1983, Barnes 2005)
  - Item response learning model (Draney, et al., 1995)
  - Item response assessment models (DiBello, et al., 1995; Embretson, 1997; von Davier, 2005)
- Machine Learning & AI
  - Combinatorial search (Russell & Norvig, 2003)

Item Labeling & the "P Matrix": Adding Alternative Factors

- How to improve existing cognitive model?
- Have experts look for *difficulty factors* that are candidates for new KCs. Put these in "P matrix"

Q Matrix

Item | Skill	Add	Sub	Mul
2*8	0	0	1
2*8 – 3	0	1	1
2*8 - 30	0	1	1
3+2*8	1	0	1

P Matrix

Item | Skill	Deal with negative	Order of Ops	…
2*8	0	0	
2*8 – 3	0	0	
2*8 - 30	1	0	
3+2*8	0	1	

Using P matrix to update Q matrix

- Create a new Q' by using elements of P as arguments to operators
  - Add operator: Q' = Q + P[,1]
  - Split operator: Q' = Q[, 2] * P[,1]

Q- Matrix after add P[, 1]

Item | Skill	Add	Sub	Mul	Div	neg
2*8	0	0	1	0	0
2*8 – 3	0	1	1	0	0
2*8 - 30	0	1	1	0	1

Q- Matrix after splitting P[, 1], Q[,2]

Item | Skill	Add	Sub	Mul	Div	Sub-neg
2*8	0	0	1	0	0
2*8 – 3	0	1	1	0	0
2*8 - 30	0	0	1	0	1

LFA: KC Model Search

- How to find best model given Q and P matrices?
- Use best-first search algorithm (Russell & Norvig 2002)
  - Guided by a heuristic, such as BIC or AIC
- Do model selection within space of Q matrices

  Steps:
  1. Start from an initial "node" in search graph using given Q
  2. Iteratively create new child nodes (Q') by applying operators with arguments from P matrix
  3. Employ heuristic (BIC of Q') to rank each node
  4. Select best node not yet expanded & go back to step 2

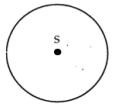
Example in Geometry of split based on factor in P matrix

Original Q matrix

Factor in P matrix

Student	Step	Skill	Opportunity	Embed
A	p1s1	Circle-area	0	alone
A	p2s1	Circle-area	1	embed
A	p2s2	Rectangle-area	0	
A	p2s3	Compose-by-add	0	
A	p3s1	Circle-area	2	alone

LAWN_SPRINKLER: SECTION FOUR, #2



Problem Statement

If a lawn sprinkler spins 360 degrees and has a 17 ft spray, find the area of lawn that will be watered if the sprinkler is not moved.

	Length of Spray (OD) - Radius	Watered Lawn - Circle
Units	feet	sq. feet
Question 1	17	907.92

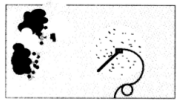
After Splitting Circle-area by *Embed*

New Q matrix

Revised Opportunity

Student	Step	Skill	Opportunity
A	p1s1	Circle-area-alone	0
A	p2s1	Circle-area-embed	0
A	p2s2	Rectangle-area	0
A	p2s3	Compose-by-add	0
A	p3s1	Circle-area-alone	1

LAWN_SPRINKLER_2: SECTION FOUR, #5



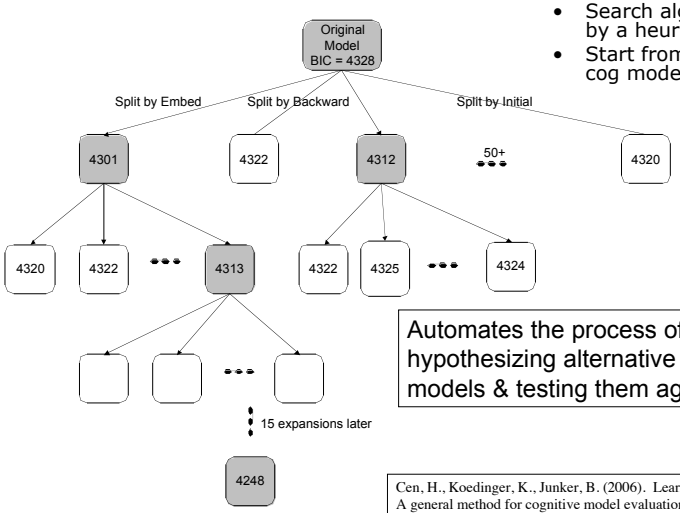
Problem Statement

A lawn sprinkler spins 360 degrees and has a 1 ft spray. If the rectangular lawn is 5 feet by 9 feet, find the area of lawn that will not be watered if the sprinkler is not moved.

	Length of Spray (OD) - (Radius)	Watered Lawn (Circle)	Area of Total Lawn (Rectangle)	Area of Unwatered Lawn
Units	cm	cm	sq. cm	cm
Question 1	1	3.1416	45	41.86

LFA –Model Search Process

- Search algorithm guided by a heuristic: BIC
- Start from an existing cog model (Q matrix)



Automates the process of hypothesizing alternative cognitive models & testing them against data

Cen, H., Koedinger, K., Junker, B. (2006). Learning Factors Analysis: A general method for cognitive model evaluation and improvement. *8th International Conference on Intelligent Tutoring Systems*.

Example LFA Results: Applying splits to original model

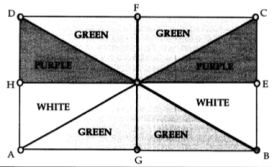
Model 1	Model 2	Model 3
Number of Splits:3	Number of Splits:3	Number of Splits:2
1. Binary split compose-by-multiplication by figurepart segment 2. Binary split circle-radius by repeat repeat 3. Binary split compose-by-addition by backward backward	1. Binary split compose-by-multiplication by figurepart segment 2. Binary split circle-radius by repeat repeat 3. Binary split compose-by-addition by figurepart area-difference	1. Binary split compose-by-multiplication by figurepart segment 2. Binary split circle-radius by repeat repeat
Number of Skills: 18	Number of Skills: 18	Number of Skills: 17
BIC: 4,248.86	BIC: 4,248.86	BIC: 4,251.07

- Common results:
  - Compose-by-multiplication split based on whether it was an area or a segment being multiplied
  - Circle-radius is split based on whether it is being done for the first time in a problem or is being repeated

Compose-by-multiplication KC examples

Composing Areas

DESIGNING_A_QUILT: SECTION TWO, #5



Problem Statement

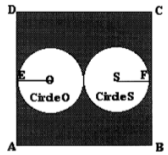
This rectangular sketch is representative of one quilt patch.You will need a total 8 patches to complete the quilt. Find the amount of each color of fabric needed to make the quilt. Note that there are four green triangles, two white triangles and two purple triangles in each patch.

1. If DF = FC = 9 inches and DH = HA = 4 inches.
2. If DF = FC = 2 inches and DH = HA = 2 inches.
3. If DF = FC = 27 inches and DH = HA = 12 inches.

	Width of Triangle	Height of Triangle	Area of Triangle	Total Area of White	Total Area of Green	Total Area of Purple
Units	inches	inches	sq. inches	sq. inches	sq. inches	sq. inches
Question 1	9	4	18	36	72	36
Question 2	2	2	2	4	8	4
Question 3	27	12	162	324	648	324

Composing Segments

TWO-CIRCLES-IN-A-SQUARE: SECTION FIVE, #6



Problem Statement

In the diagram, two identical circles, O and S, are enclosed in Square ABCD. Circle O has a radius (OE) of 1.0 feet. Circle S has a radius (SF) of identical length to Circle O.

Using this information, find the area of the shaded region.

	Length of radius (OE)	Length (AB) of square's side	Area of shaded region	Area of square ABCD	Area of unshaded region
Units	inches	inches	sq. inches	sq. inches	sq. inches
Question 1	1	4	9.72	16	6.28

Tutor Design Implications 1

- LFA search suggests distinctions to address in instruction & assessment

  With these new distinctions, tutor can
  - Generate hints better directed to specific student difficulties
  - Improve knowledge tracing & problem selection for better cognitive mastery
- Example: Consider Compose-by-multiplication before LFA

	Intercept	slope	Avg Practice Opportunties	Initial Probability	Avg Probability	Final Probability
CM	-.15	.1	10.2	.65	.84	.92

With final probability .92, many students are short of .95 mastery threshold

Tutor Design Implications 2

- However, after split:

	Intercept	slope	Avg Practice Opportunties	Initial Probability	Avg Probability	Final Probability
CM	-.15	.1	10.2	.65	.84	.92
CMarea	-.009	.17	9	.64	.86	.96
CMsegment	-1.42	.48	1.9	.32	.54	.60

- CM-area and CM-segment look quite different
  - CM-area is now above .95 mastery threshold (at .96)
  - But CM-segment is only at .60
  - Original model penalizes students who have key idea about composite areas (CM-area) -- some students solve more problems than needed
- Instructional redesign implications:
  - *Change skillometer* so CM-area & CM-segment are separately addressed
    - Set parameters appropriately -- CM-segment with have a lower initial known value
  - *Add more problems* to allow for mastery of CM-segment
  - *Add new hints* specific to the CM-segment situation

Summary of Learning Factors Analysis (LFA)

- LFA combines statistics, human expertise, & combinatorial search to discover cognitive models
- Evaluates a single model in seconds, searches 100s of models in hours
  - Model statistics are meaningful
  - Improved models suggest tutor improvements
- Can currently be applied, by request, to any dataset in DataShop with at least two KC models

Mixed initiative human-machine discovery

1. Human
   - Hypothesize possible “learning factors” and code steps
2. Machine
   - Search over factors, report best models discovered
3. Human
   - Inspect results
   - If needed, propose new factors. Go to 2.
   - If good, modify tutor and test.

Human-machine discovery of new cognitive models

- Better models discovered in Geometry, Statistics, English, Physics

Dataset name	Domain	Existing best		LFA Discovered		Improved
		BIC	# of KCs	BIC	# of KCs	BIC
Geometry96-97	Geometry Area	5606	12	5548	10	1%
Hampton0506	Geometry Area	15047	18	12476	15	17%
Cog discovery	Geometry Area	31183	49	31109	29	0.2%
Statistics - Fall 2009	Statistics	3611	14	3454	8	4%
IWT S-E Study 2	English articles	7162	19	7068	11	1%
Physics 2009 Spring	Physics	27051	239	24917	14	8%

Some Open EDM Research Problems

Open Research Questions: Technical

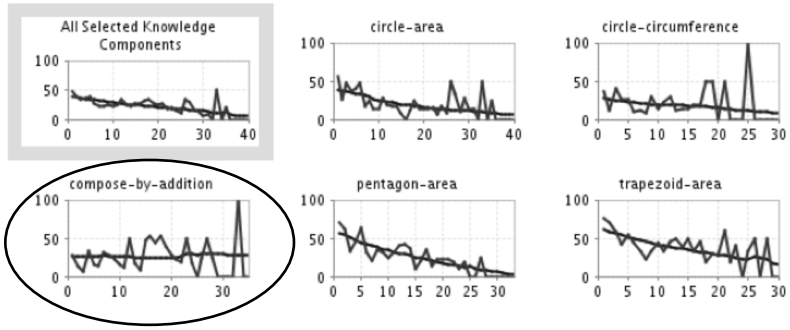
- What factors to consider? P matrix is hard to create
  - Enhancing human role: Data visualization strategies
  - Other techniques: Matrix factorization, LiFT
  - Other data: Do clustering on problem text
- Interpreting LFA output can be difficult
  - How to make interpretation easier?

=> Researcher can’t just “go by the numbers”
1) Understand the domain, the tasks
2) Get close to the data

Model search using DataShop:
Human & machine improvements

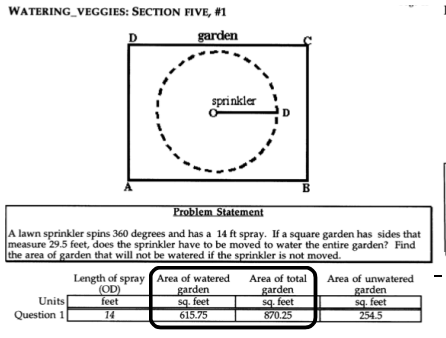
- DataShop datasets w/ improved KC models:
  - Geometry Area (1996-1997), Geometry Area Hampton 2005-2006 Unit 34, …
- New KCs (learning factors) found using DataShop visualization tools
  - Learning curve, point tool, performance profiler
  - Example of human "feature engineering"
- New KC models also discovered by LFA
  - Research goal: Iterate between LFA & visualization to find increasingly better KC models

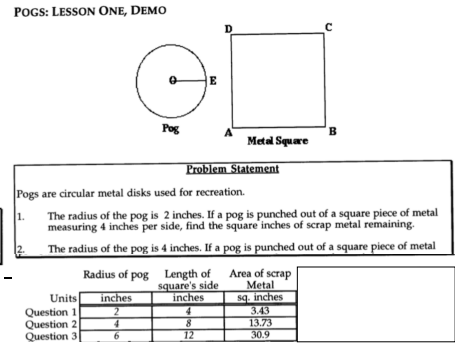
Most curves "curve", but if flat, then KC may be bad



Detecting planning skills:
Scaffolded vs. unscaffolded problems

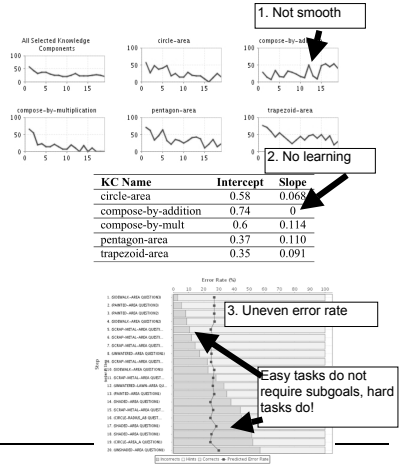
- Scaffolded
  - Prompts are given for subgoals
- Unscaffolded
  - Prompts are not given for subgoals (initially)



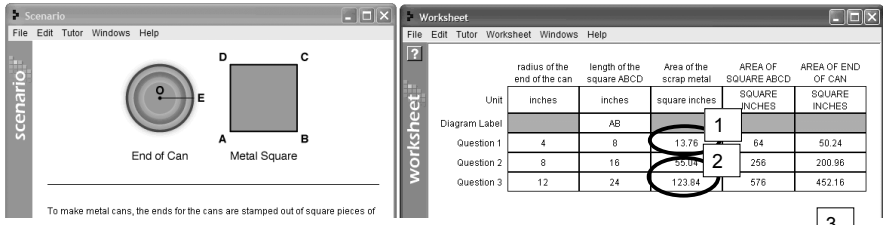


Discovering a new knowledge component

- Each KC should have:
  1. smooth learning curve
  2. statistical evidence of learning
  3. even error rates across tasks
- Create new KCs by finding a feature common to hard tasks but missing in easy ones



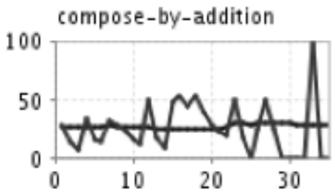
New model discovery: Split "compose" into 3 skills



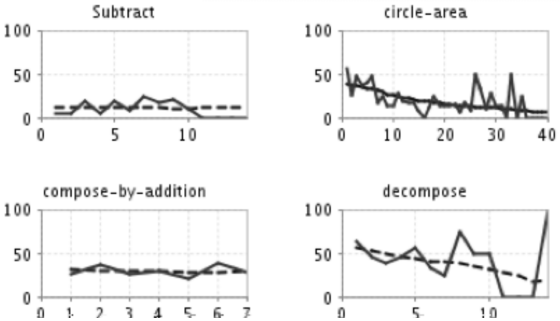
- Hidden planning knowledge:
  If you need to find the area of an irregular shape, **then** try to find the areas of regular shapes that make it up
- Redesign instruction in tutor
  - Design tasks that isolate the hidden planning skill
  - Given square & circle area, find leftover

When prompts are initially present for component areas

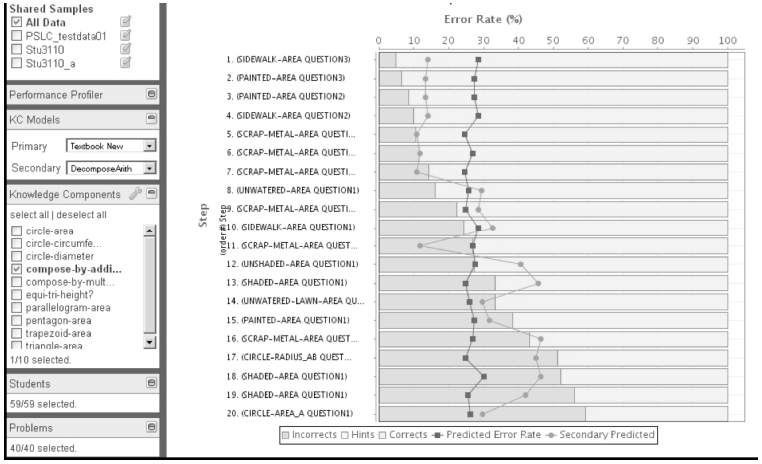
Before unpacking compose-by-addition



After -- unpacked into subtract, decompose, remaining compose-by-addition



3-way split in new model (green) better fits variability in error rates than original (blue)



Automate human-machine strategies for "blip" detection

- Research goal: Automate low slope, non-low intercept, & high residual detection
- Uses:
  - speed up LFA search
  - point human coders to bad KCs
    - cluster harder vs. easier tasks

Developing & evaluating different learning curve models

- Many papers in Educational Data Mining (EDM) conference
  - Also in Knowledge Discovery & Data mining (KDD)
- Papers comparing knowledge tracing, AFM, PFA, CPFA, IFA
  - See papers by Pavlik, Beck, Chi …

Open Research Questions: Psychology of Learning

- Change AFM model assumptions
  - Is student learning rate really constant?
    - Does a Student x Opportunity interaction term improve fit?
    - What instructional conditions or student factors change rate?
  - Is knowledge space "uni-dimensional"?
    - Does a Student x KC interaction term improve fit?
  - Need different KC models for different students/conditions?
- Is learning curve an exponential or power law?
  - Long-standing debate, which has focused on "reaction time" not on error rate!
  - Compare use of Opportunity vs.Log(Opportunity)
- Other outcome variables: reaction time, assistance score
- Other predictors: Opportunities => Time per instructional event; Kinds of opportunities: Successes, failures, hints, gamed steps, …

Open Research Questions: Instructional Improvement

- Do LFA results generalize across data sets?
  - Is AIC or BIC a good estimate for cross-validation results?
  - Does a model discovered with one year's tutor data generalize to a next year?
  - Does model discovery work for ed games, other domains?
- Use learning curves to compare instructional conditions in experiments
- *Need more "close the loop" experiments*
  - EDM => better model => better tutor => better student learning