# 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.

Cen, H., Koedinger, K., Junker, B. **Is Over Practice Necessary? Improving Learning Efficiency with the Cognitive Tutor**. *13th International Conference on Artificial Intelligence in Education*. 2007.

# Domain-Specific Cognitive Models

- Question: How do students represent knowledge in a given domain?
- Answering this question involves deep domain analysis
- The product is a cognitive model of students' knowledge
  - Recall cognitive models drive ITS behaviors & instructional design decisions

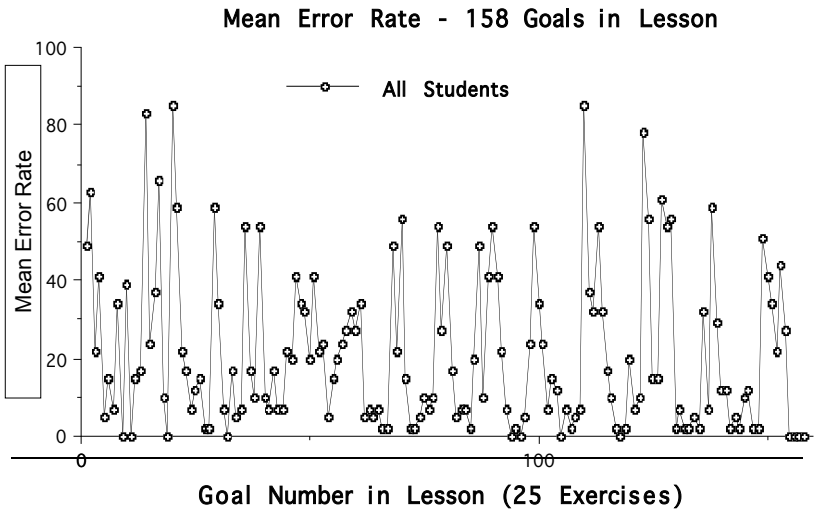
# Knowledge Decomposibility Hypothesis

- Human acquisition of academic competencies can be decomposed into units, called knowledge components (KCs), that *predict student task performance & transfer*
- *Performance* predictions
  - If item I1 only requires KC1 & item I2 requires both KC1 and KC2, then item I2 will be harder than I1
  - If student can do I2, then they can do I1
- *Transfer* predictions
  - If item I1 requires KC1, & item I3 also requires KC1, then practice on I3 will improve I1
  - If item I1 requires KC1, & item I4 requires only KC3, then practice on I4 will ***not*** improve I1
- Fundamental EDM idea:
  - *We can discover KCs (cog models) by working these predictions backwards!*

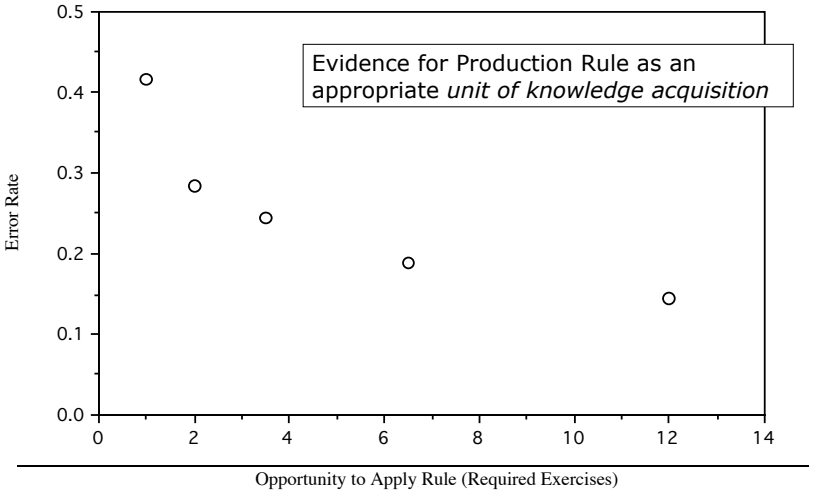
**Example of Items & KCs**

| | KC1 add | KC2 carry | KC3 subt |
|---|---|---|---|
| I1: 5+3 | 1 | 0 | 0 |
| I2: 15+7 | 1 | 1 | 0 |
| I3: 4+2 | 1 | 0 | 0 |
| I4: 5-3 | 0 | 0 | 1 |

# Student Performance As They Practice with the LISP Tutor



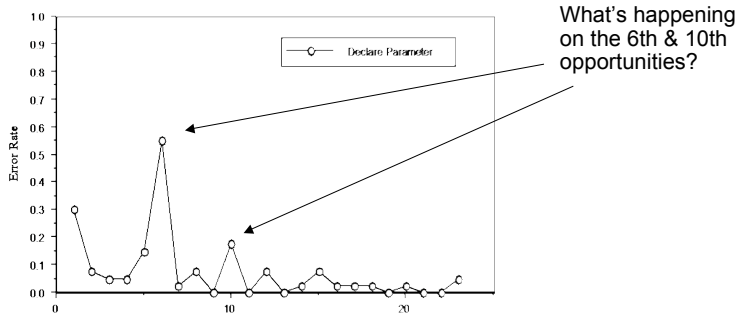
# Production Rule Analysis




# 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 …

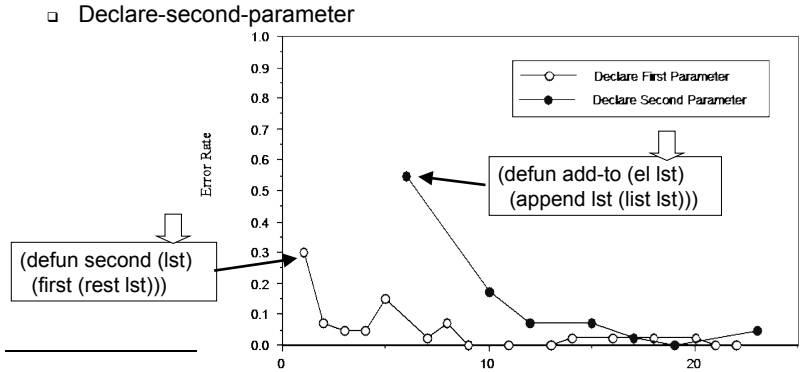
# Curve for "Declare Parameter" production rule




- 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?

- Learning curve analysis
  - Identify blips by hand & eye
  - Manually create a new model
  - Qualitative judgment
- Need to automatically:
  - Identify blips by system
  - Propose alternative cognitive models
  - Evaluate each model quantitatively

# Learning Factors Analysis

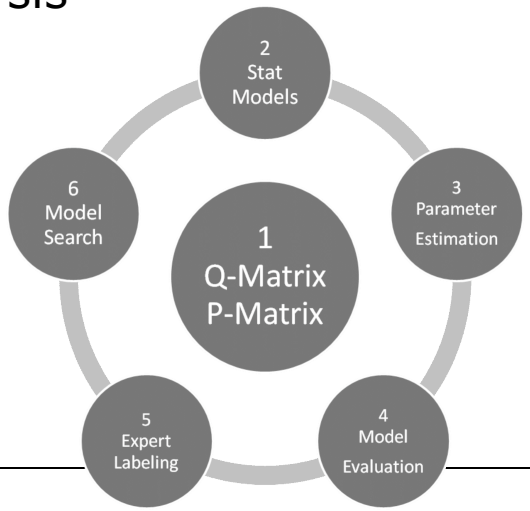
## Learning Factors Analysis (LFA): A Tool for KC Analysis

- LFA is a method for discovering & evaluating alternative cognitive models
  - Finds knowledge component decomposition that best predicts student performance & learning transfer
- Inputs
  - Data: Student success on tasks in domain over time
  - Codes: Factors hypothesized to drive task difficulty & transfer
    - A mapping between these factors & domain tasks
- Outputs
  - A rank ordering of most predictive cognitive models
  - For each model, a measure of its generalizability & parameter estimates for knowledge component difficulty, learning rates, & student proficiency

## Learning Factors Analysis (LFA) draws from multiple disciplines

- Machine Learning & AI
  - Combinatorial search (Russell & Norvig, 2003)
  - Exponential-family principal component analysis (Gordon, 2002)
- 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)
- Cognitive Psychology
  - Learning curve analysis (Corbett, et al 1995)

# Steps in Learning Factors Analysis




# Representing Knowledge Components as factors of items

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

Items X Knowledge Components (KCs)

| Item \| Skills: | 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
  - 2*8 above
- Multi-KC item = when a row has many 1's
  - 2*8 – 3

What good is a Q matrix? Can predict student accuracy *on items not previously seen*, based on KCs involved

# The Statistical Model

- Generalized Power Law to fit learning curves
  - Logistic regression (Draney, Wilson, Pirolli, 1995)
- Assumptions
  - Some skills may easier from the start than others
    => use an *intercept* parameter for each skill
  - Some skills are easier to learn than others
    => use a *slope* parameter for each skill
  - Different students may initially know more or less
    => use an *intercept* parameter for each student
  - Students generally learn at the same rate
    => no *slope* parameters for each student

Prior Summer School project!

- These assumptions are reflected in a statistical model …

# Simple Statistical Model of Performance & Learning

- Problem: How to predict student responses from model?
- Solutions: Additive Factor Model (Draney, et al. 1995)

$$p_{ij} = \Pr(Y_{ij} = 1 \mid \theta_i, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{\exp(\theta_i + \sum_{k=1}^{K} q_{jk}\beta_k + \sum_{k=1}^{K} q_{jk}\gamma_k T_{ik})}{1 + \exp(\theta_i + \sum_{k=1}^{K} q_{jk}\beta_k + \sum_{k=1}^{K} q_{jk}\gamma_k T_{ik})}$$

$$q_{jk} = \begin{cases} 1 & \text{item j uses skill k} \\ 0 & \text{otherwise} \end{cases}$$

$Y_{ij}$ = the response of student i on item j
$\theta_i$ = coefficient for student i
$\beta_k$ = coefficient for skill k
$\gamma_k$ = coefficient for the learning rate of skill k
$T_{ik}$ = the number of practice opportunities student i has had on the skill k

## Comparing Additive Factor Model to other psychometric techniques

- Instance of generalized linear regression, binomial family or "logistic regression"
  - R code: glm(success~student+skill+skill:opportunity, family=binomial,…)
- Extension of item response theory
  - IRT has simply a student term (theta-i) + item term (beta-j)
  - R code: glm(success~student+item, family=binomial,…)
  - The additive factor model behind LFA is different because:
    - It breaks items down in terms of knowledge component factors
    - It adds term for practice opportunities per component

$$p_{ij} = \Pr(Y_{ij} = 1 \mid \theta_i, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{\exp(\theta_i + \sum_{k=1}^{K} q_{jk}\beta_k + \sum_{k=1}^{K} q_{jk}\gamma_k T_{ik})}{1 + \exp(\theta_i + \sum_{k=1}^{K} q_{jk}\beta_k + \sum_{k=1}^{K} q_{jk}\gamma_k T_{ik})}$$

$Y_{ij}$ = the response of student i on item j
$\theta_i$ = coefficient for student i
$\beta_k$ = coefficient for skill k

$$q_{jk} = \begin{cases} 1 & \text{item j uses skill k} \\ 0 & \text{otherwise} \end{cases}$$

## Model Evaluation

- How to compare cognitive models?
  - A good model minimizes prediction risk by balancing fit with data & complexity (Wasserman 2005)
- Compare BIC for the cognitive models
  - BIC is "Bayesian Information Criteria"
  - BIC = -2*log-likelihood + numPar * log(numOb)
  - Better (lower) BIC == better predict data that haven't seen
- Mimics cross validation, but is faster to compute



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

- Problem: How to improve existing cognitive model?
- Solution: 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

- Problem: How to find best model given Q and P matrices?
- Solution: Combinatorial search

- A best-first search algorithm (Russell & Norvig 2002)
  - Guided by a heuristic, such as BIC
- Goal: Do model selection within logistic regression model space

  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

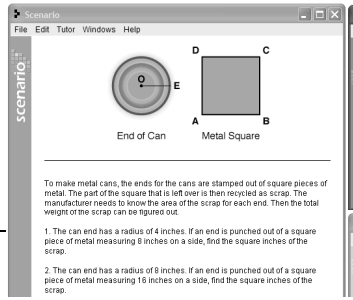
# Learning Factors Analysis: Example in Geometry Area

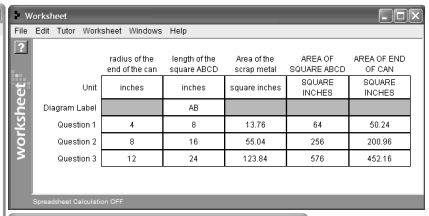
# Area Unit of Geometry Cognitive Tutor

- 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 LFA

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) | Skill (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

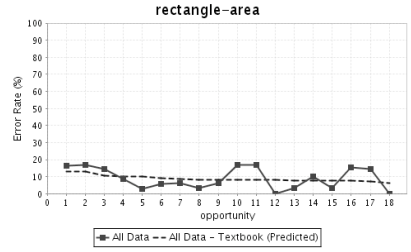
| Model Statistics | |
|---|---|
| AIC | 3,950 |
| BIC | 4,285 |
| MAD | 0.083 |

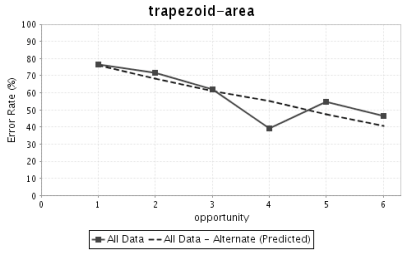
The AIC, BIC & MAD statistics provide alternative ways to evaluate models

MAD = Mean Absolute Deviation

# Application: Use Statistical Model to improve tutor

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



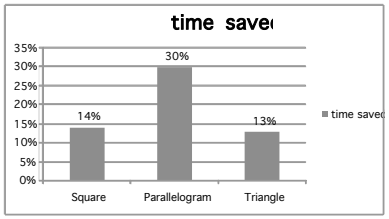

initial error rate 12%
reduced to 8%
after 18 times of practice

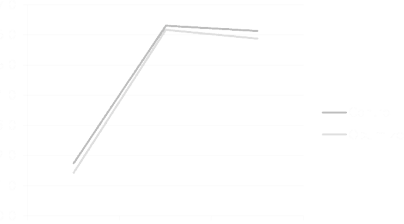
initial error rate 76%
reduced to 40%
after 6 times of practice



# "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





# 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

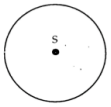
After Splitting Circle-area by *Embed*

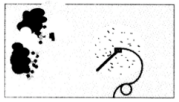
New Q matrix

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**

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 |

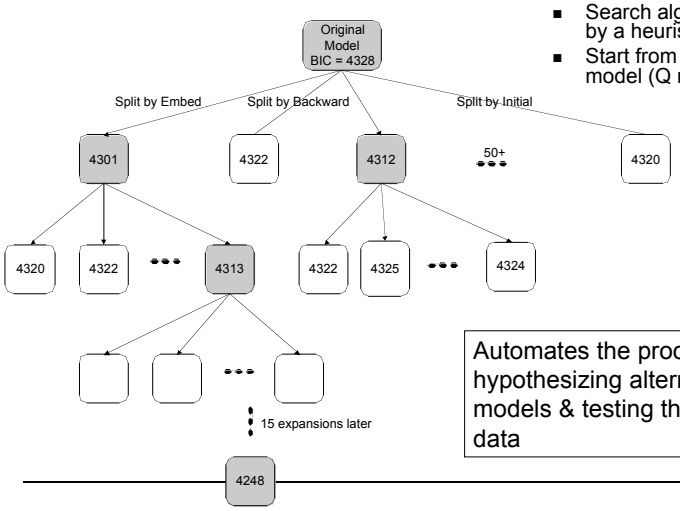
**Problem Statement**

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 KC model (Q matrix)




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

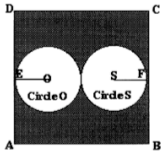
# LFA Results 1: Applying splits to original model

| 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<br>2. Binary split circle-radius by repeat repeat<br>3. Binary split compose-by-addition by backward backward | 1. Binary split compose-by-multiplication by figurepart segment<br>2. Binary split circle-radius by repeat repeat<br>3. Binary split compose-by-addition by figurepart area-difference | 1. Binary split compose-by-multiplication by figurepart segment<br>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
- Made sense, but less than expected …

# Other Geometry problem examples

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



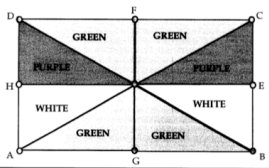

**Problem Statement**

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 |

DESIGNING_A_QUILT: SECTION TWO, #5

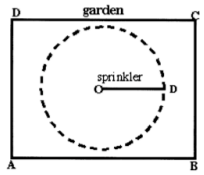



**Problem Statement**

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 |




**Problem Statement**

A lawn sprinkler spins 360 degrees and has a 14 ft spray. If a square garden has sides that measure 29.5 feet, does the sprinkler have to be moved to water the entire garden? Find the area of garden that will not be watered if the sprinkler is not moved.

| | Length of spray (OD) | Area of watered garden | Area of total garden | Area of unwatered garden |
|---|---|---|---|---|
| Units | feet | sq. feet | sq. feet | sq. feet |
| Question 1 | 14 | 615.75 | 870.25 | 254.5 |

# Example of Tutor Design Implications

- 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

# Making a distinction changes assessment decision

- 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
- Implications:
  - Original model penalizes students who have key idea about composite areas (CM-area) -- some students solve more problems than needed
  - CM-segment is not getting enough practice
    - Instructional design choice: Add instructional objective & more problems or not?

# Perhaps original model is good enough -- Can LFA recover it?

- Merge some skills in original model, to produce 8 skills:
  - Circle-area, Circle-radius => Circle
  - Circle-circumference, Circle-diameter => Circle-CD
  - Parallelogram-area, Parallelogram-side => Parallelogram
  - Pentagon-area, Pentagon-side => Pentagon
  - Trapezoid-area, Trapezoid-base, Trapezoid-height => Trapezoid
  - Triangle-area, Triangle-side => Triangle
  - Compose-by-addition
  - Compose-by-multiplication
- Does splitting by “backward” (or otherwise) yield a better model? Closer to original?

# LFA Results 2: Recovery

| Model 1 | Model 2 | Model 3 |
|---|---|---|
| Number of Splits: 4 | Number of Splits: 3 | Number of Splits: 4 |
| Circle*area<br>Circle*radius*initial<br>Circle*radius*repeat<br>Compose-by-addition<br>Compose-by-addition*area-difference<br>Compose-by-multiplication*area-combination<br>Compose-by-multiplication*segment | All skills are the same as those in model 1 except that<br>1. Circle is split into Circle *backward*initial, Circle *backward*repeat, Circle*forward,<br>2. Compose-by-addition is not split | All skills are the same as those in model 1 except that<br>1. Circle is split into Circle *backward*initial, Circle *backward*repeat, Circle *forward<br>2. Compose-by-addition is split into Compose-by-addition and Compose-by-addition*segment |
| Number of skills: 12 | Number of skills: 11 | Number of skills: 12 |
| BIC: 4,169.315 | BIC: 4,171.523 | BIC: 4,171.786 |

- Only 1 recovery: Circle-area vs. Circle-radius
- More merged model fits better
  - Why? More transfer going on than expected or not enough data to make distinctions? Other relevant data sets …

# Research Issues & Summary

## Open Research Questions: Technical

- What factors to consider? P matrix is hard to create
  - Enhancing human role: Data visualization strategies
  - Other techniques: Principal Component Analysis +
  - Other data: Do clustering on problem text
- Interpreting LFA output can be difficult
  - LFA outputs many models with roughly equivalent BICs
  - How to select from large equivalence class of models?
  - How to interpret results?

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

## DataShop Case Study video

- "Using DataShop to discover a better knowledge component model of student learning"

## 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
- Other applications of LFA & model comparison
- Used by others:
  - Individual differences in learning rate (Rafferty et. al., 2007)
  - Alternative methods for error attribution (Nwaigwe, et al. 2007)
  - Model comparison for DFA data in math (Baker; Rittle-Johnson)
  - Learning transfer in reading (Leszczenski & Beck, 2007)

## Open Research Questions: Psychology of Learning

- Test statistical model assumptions: Right terms?
  - 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?
- Right shape: Power law or an exponential?
  - Long-standing hot debate
  - Has focused on "reaction time" not on error rate!
- Other predictor & outcome variables (x & y of curve)
  - Outcome: Error rate => Reaction time, assistance score
  - Predictor: Opportunities => Time per instructional event

## Open Research Questions: Instructional Improvement

- Do LFA results generalize across data sets?
  - Is 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 in ill-structured domains?
- Use learning curves to compare instructional conditions in experiments
- *Need more "close the loop" experiments*
  - EDM => better model => better tutor => better student learning

## END

## Overview

- Learning Factors Analysis algorithm
- A Geometry Cognitive Model and Log Data
  - Experiments and Results
  - DataShop case study video
- Learning curves between conditions

## Future Goals in Discovering Domain Models

- Improve knowledge-discovery methods
  - Partial Order Knowledge Structures (POKS)
  - Exponential-family Principle Component Analysis
- Improve human-machine interaction
  - Better process for task difficulty factor labeling
- Show models yield improved student learning

## Lots of interesting questions to be addressed with Ed Data Mining!!

- Assessment questions
  - Can on-line embedded assessment replace standardized tests?
  - Can assessment be accurate if students are learning during test?
- Learning theory questions
  - What are the "elements of transfer" in human learning?
  - Is learning *rate* driven by student variability or content variability?
  - Can conceptual change be tracked & better understood?
- Instructional questions
  - What instructional moves yield the greatest increases in learning?
  - Can we replace ANOVA with learning curve comparison to better evaluate learning experiments?
- Metacogniton & motivation questions
  - Can student affect & motivation be detected in on-line click stream data?
  - Can student metacognitive & self-regulated learning strategies be detected in on-line click stream data?

## EXTRA SLIDES

Past SS Project
Physics curves
Condition contrast

## Past Summer School Project

Rafferty (Stanford) & Yudelson (U Pitt)

## Past Project Example

- Rafferty (Stanford) & Yudelson (Pitt)
- Analyzed a data set from Geometry
- Applied Learning Factors Analysis (LFA)
- Driving questions:
  - Are students learning at the same rate as assumed in prior LFA models?
  - Do we need different cognitive models (KC models) to account for low-achieving vs. high-achieving students?

## A Statistical Model for Learning Curves

- Predicts whether student is correct depending on knowledge & practice
- Additive Factor Model (Draney, et al. 1995, Cen, Koedinger, Junker, 2006)

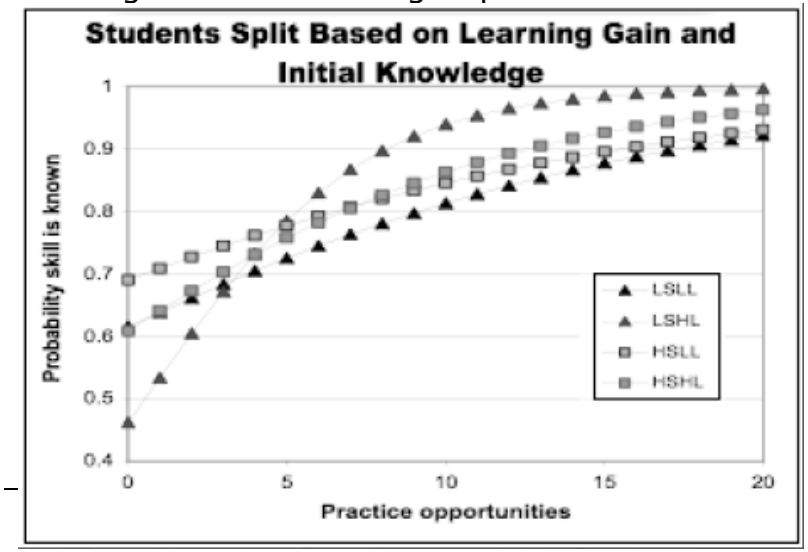
$$p_{ij} = \Pr(Y_{ij} = 1 \mid \theta_i, \boldsymbol{\beta}, \gamma) = \frac{\exp(\theta_i + \sum_{k=1}^{K} q_{jk}\beta_k + \sum_{k=1}^{K} q_{jk}\gamma_k T_{ik})}{1 + \exp(\theta_i + \sum_{k=1}^{K} q_{jk}\beta_k + \sum_{k=1}^{K} q_{jk}\gamma_k T_{ik})}$$

$$q_{jk} = \begin{cases} 1 & \text{item j uses skill k} \\ 0 & \text{otherwise} \end{cases}$$

$Y_{ij}$ = the response of student i on item j
$\theta_i$ = coefficient for student i
$\beta_k$ = coefficient for skill k
$\gamma_k$ = coefficient for the learning rate of skill k
$T_{ik}$ = the number of practice opportunities student i has had on the skill k

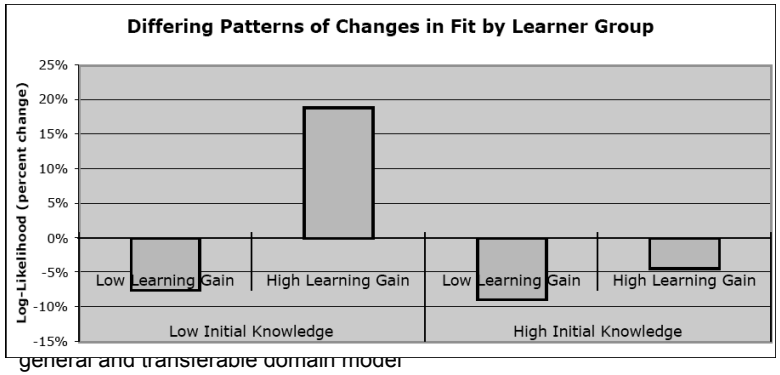
Learning rate is different for different skills, but not for different students

## Low-Start High-Learn (LSHL) group has a faster learning rate than other groups of students



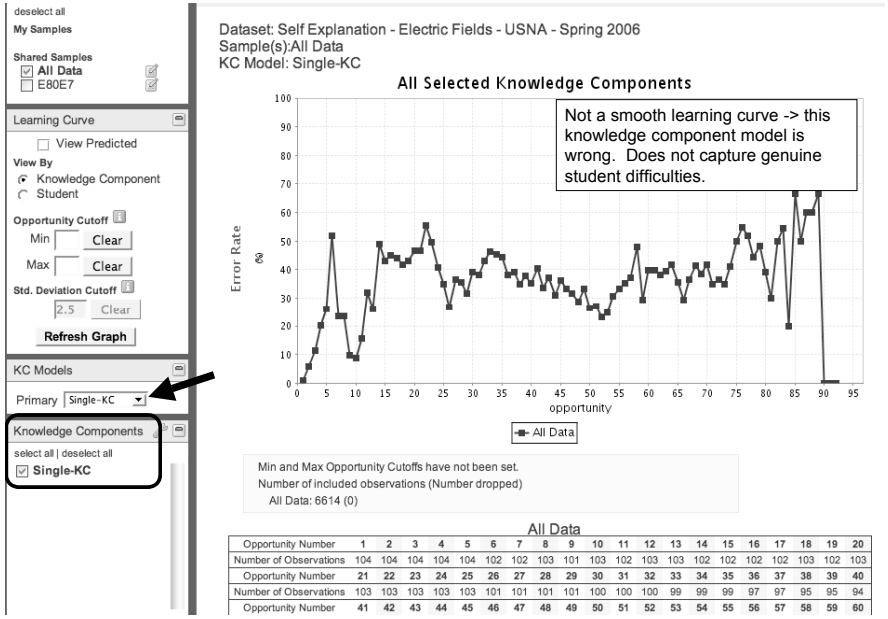

## Rafferty & Yudelson Results 2

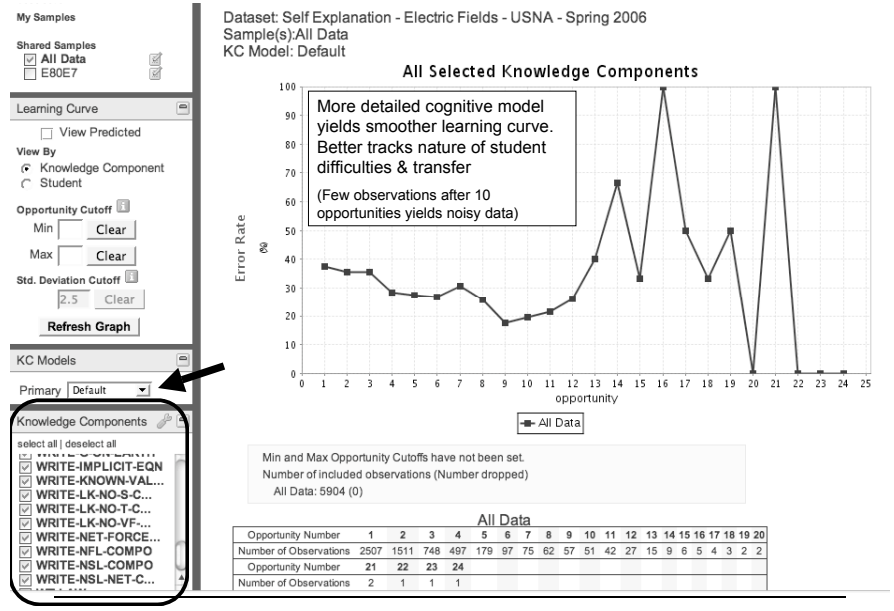
- Is it "faster" learning or "different" learning?
  - Fit with a more *compact model* is better for low start high learn

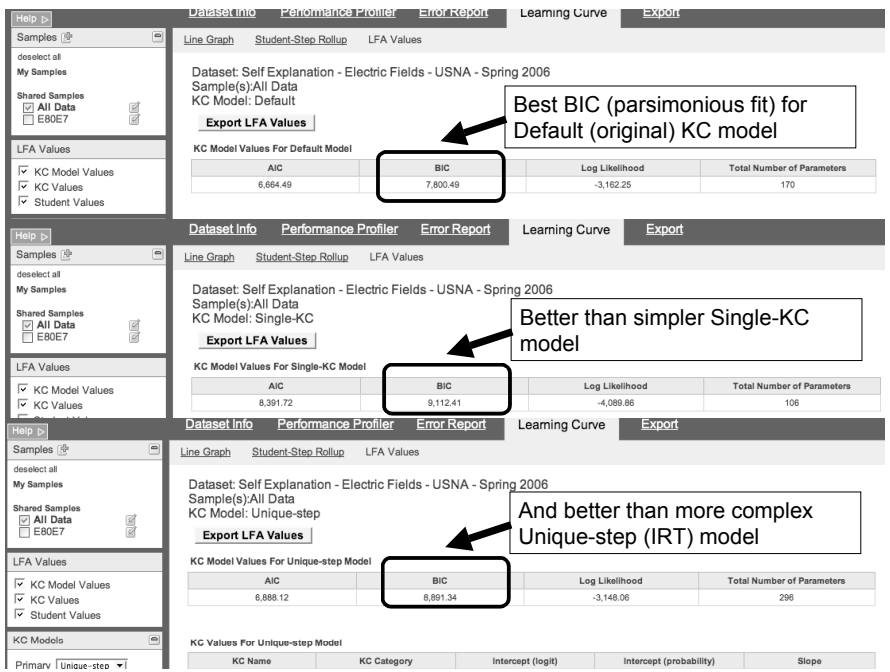



  - general and transferable domain model
- Resulted in best Young Researcher Track paper at AIED07

## Learning curve constrast in Physics dataset …




# The Long-Standing Transfer Debate

- General: Faculty theory of mind
  - Mind has faculties that can be exercised with Latin, Geometry, … video games, n-back task*

  => Transfer is broad & general, across domains
- Specific: Thorndike's identical elements
  - Mind is made up of stimulus-response elements
    - Transfer occurs between tasks with common elements
    - 1922 study: "Multiply $x^a$ & $x^b$" fails to transfer to "multiply $4^a$ & $4^b$"

  => Transfer is narrow, within domains & tasks
- More recent view
  - Singley & Anderson: Learning & transfer occurs at the grain size of the production rule
    - But begs the question: What's the grain size of a production?

*Jaeggi, Buschkuehl, Jonides, & Perrig (2008). Improving fluid intelligence with training on working memory.

# Learning curve analysis in an experiment

- Experiments on Expert vs. Intelligent Novice feedback in an Excel Programming Tutor

# Spreadsheet formulas: *Absolute referencing* stalls many learners

| | A | B |
|---|---|---|
| 1 | | Hourly Wage |
| 2 | Hours Worked | $10 |
| 3 | 20 | =A3*B2 |
| 4 | 30 | =A4*B3 |
| 5 | 40 | =A5*B4 |

(a)

| | A | B |
|---|---|---|
| 1 | | Hourly Wage |
| 2 | Hours Worked | $10 |
| 3 | 20 | $200 |
| 4 | 30 | 6000 |
| 5 | 40 | 240000 |

(b) ✗

| | A | B |
|---|---|---|
| 1 | | Hourly Wage |
| 2 | Hours Worked | $10 |
| 3 | 20 | =A3*B$2 |
| 4 | 30 | =A4*B$2 |
| 5 | 40 | =A5*B$2 |

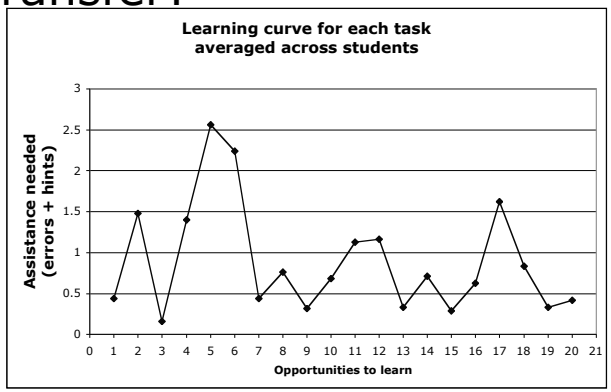
(c)

| | A | B |
|---|---|---|
| 1 | | Hourly Wage |
| 2 | Hours Worked | $10 |
| 3 | 20 | $200 |
| 4 | 30 | $300 |
| 5 | 40 | $400 |

(d) ✓

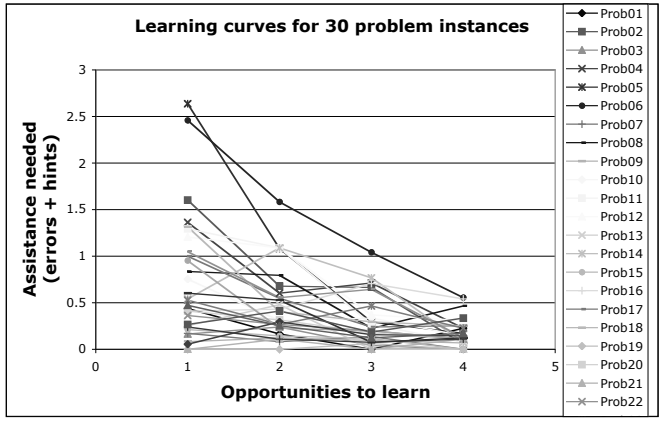
Theoretical interest: Difficult domain reveals edges of human learning

# Faculties of mind theory of transfer?



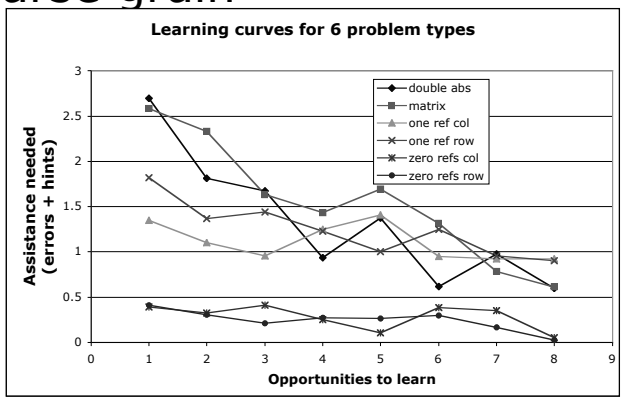

- Exercising "reasoning faculty" or "programming faculty" does *not (*clearly) reveal improvement

# Identical elements theory of transfer?



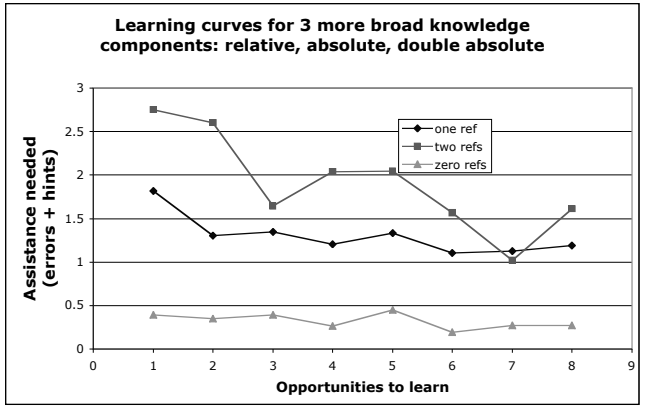

- Maybe, but perhaps too fine
- Are there groups of items across which we see transfer?

# Pattern emerges at more coarse grain




- Knowledge acquired in some problems transfer to others
- What is that knowledge?

# Broader rules also plausible




- Which one is right?
- Can we empirically compare knowledge models?

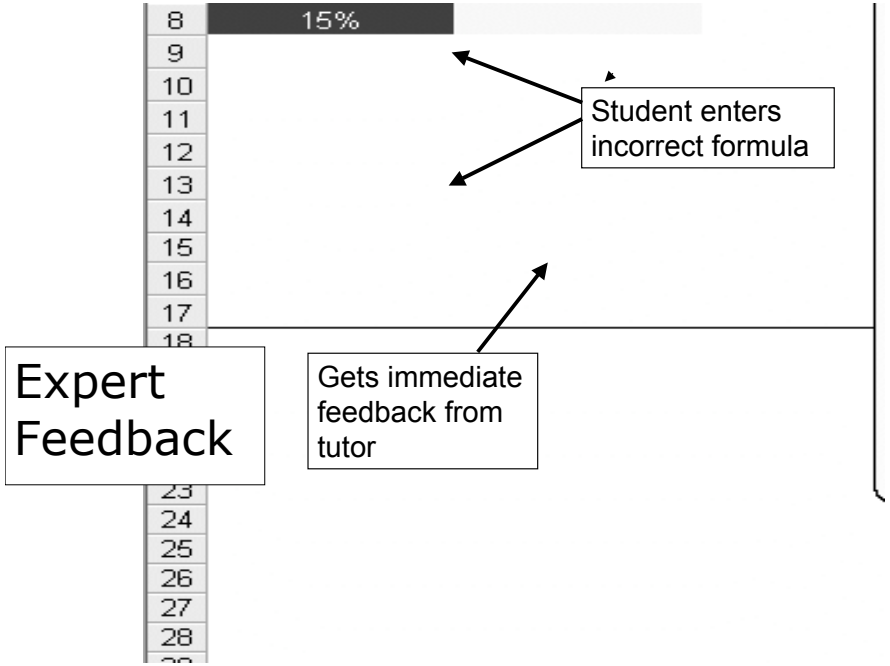
# Which knowledge model is correct?

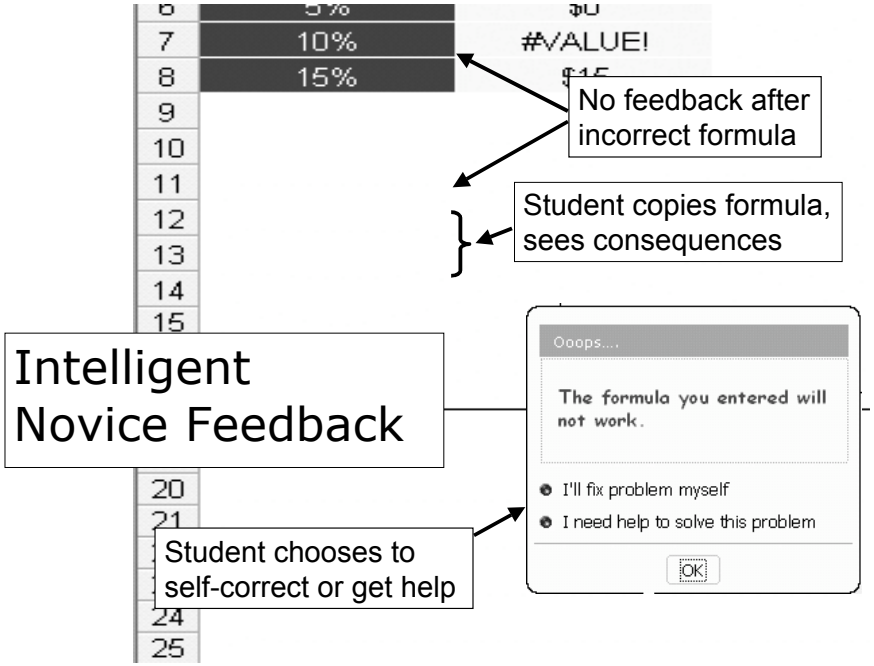
| Concepts | Model parameters | Log Likelihood | AIC | BIC |
|---|---|---|---|---|
| 30 (identical elements) | 108 | *-2070* | 4356 | 5062 |
| 6 | 60 | -2097 | *4314* | 4706 |
| 3 | 54 | -2103 | *4313* | *4666* |
| 1 (faculty theory) | *50* | -2447 | 4994 | 5321 |

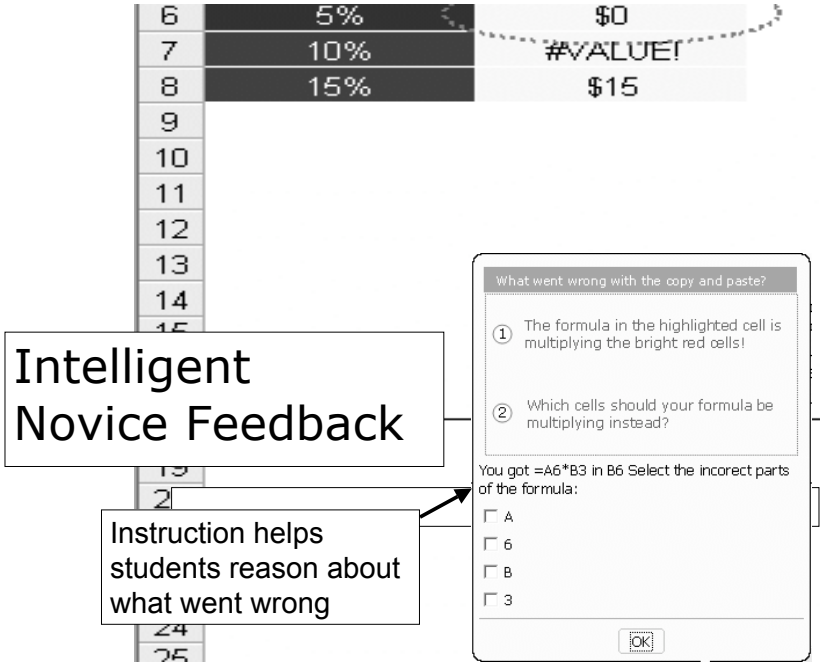
- Intermediate models are best …
- Learning curve "smoothness" selects for better cognitive models of learning

# Experimental Design:
## Intelligent Novice vs. Expert Model

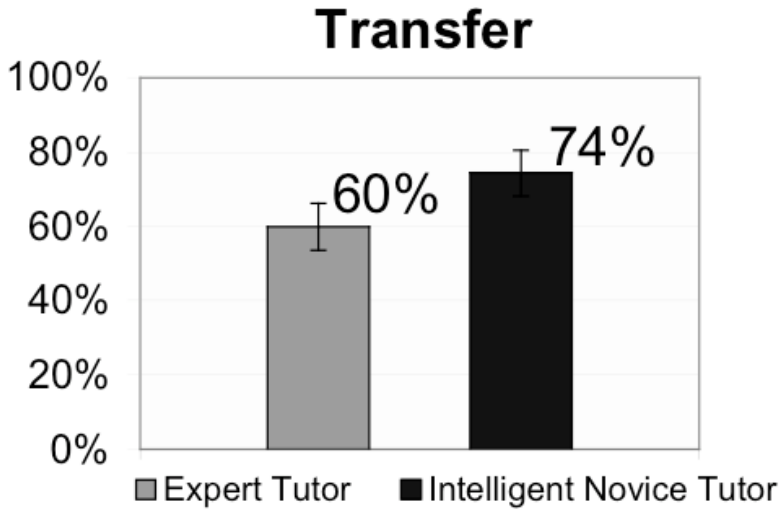
- Recast delayed vs. immediate feedback debate as contrasting "*model of desired performance*"
- Control: *Expert* Model & Tutor
  - Goal: students should not make errors
  - Feedback is *immediate* on all errors
- Treatment: *Intelligent Novice* Model & Tutor
  - Goal: students can make some errors, but should recognize & self-correct
  - Feedback appears delayed, but is *immediate* -> when student fails to self-correct own errors




## Results: IN better on all measures




All differences significant: $F(1,35)>13$, $p<0.001$

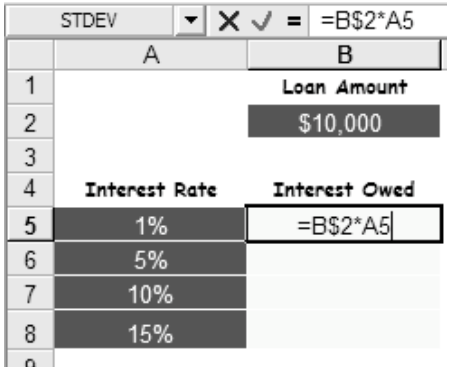
# Can we detect a qualitative difference from learning curves?

Hypotheses:

- Shallow rule model better accounts for EX student performance
- Deep rule model better accounts for IN student performance
  - Supporting error self-correction promotes deeper feature encoding

    => more general knowledge
    => greater transfer

# Shallow vs. deep learning: Depth of encoding



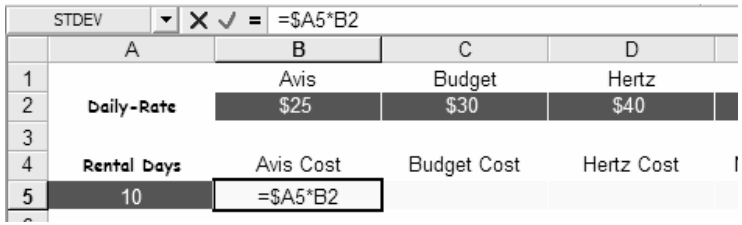

- Shallow feature encoding*:

  *Rule S1*: If you are copying down a *column*, then put a $ in front of the *number*.
- Deep feature encoding:

  *Rule D*: If a cell index needs to stay fixed but would change when copied, then put a $ in front of that index.

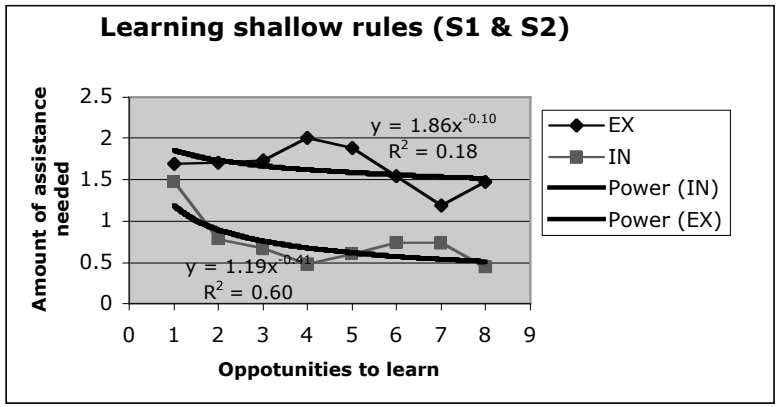
*Note: These *describe mental* structures that students *may not* be able to verbalize.

# Which knowledge component transfers to new situation?



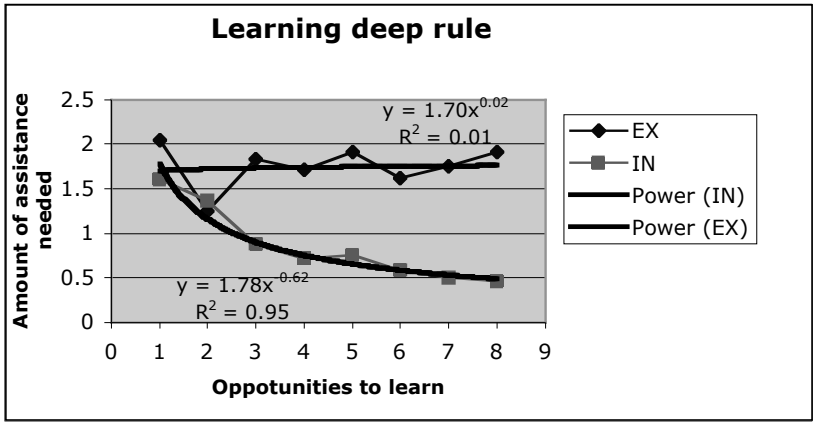

- Deep transfers
  - Index A would change when copied …
- Shallow does not
  - S1 applies to a number when copying down a column
- A new shallow rule might be learned
  - *Rule S2*: If you are copying across a row, then put a $ in front of the letter.

# Shallow rule model provides an OK fit for two groups




- IN group generally less erratic
- Better fit here for EX than in deep model …

# Deep model: Fit gets better for IN, worse for EX!




- Error self-correction support (IN) leads to deeper feature encoding

# Intelligent Novice Tutor => More General Knowledge

- Beyond assessments of ability or achievement
  - Assess abilities during learning
  - Crease assessments of knowledge
- Beyond single test score
  - What do students know
  - What is range of their transfer