A Framework for Multifaceted Evaluation of Student Models

Yun Huang[1] José P. González-Brenes[2]
Rohit Kumar[3] Peter Brusilovsky[1]

[1]University of Pittsburgh
[2]Pearson Research & Innovation Network
[3]Speech, Language and Multimedia Raytheon BBN Technologies

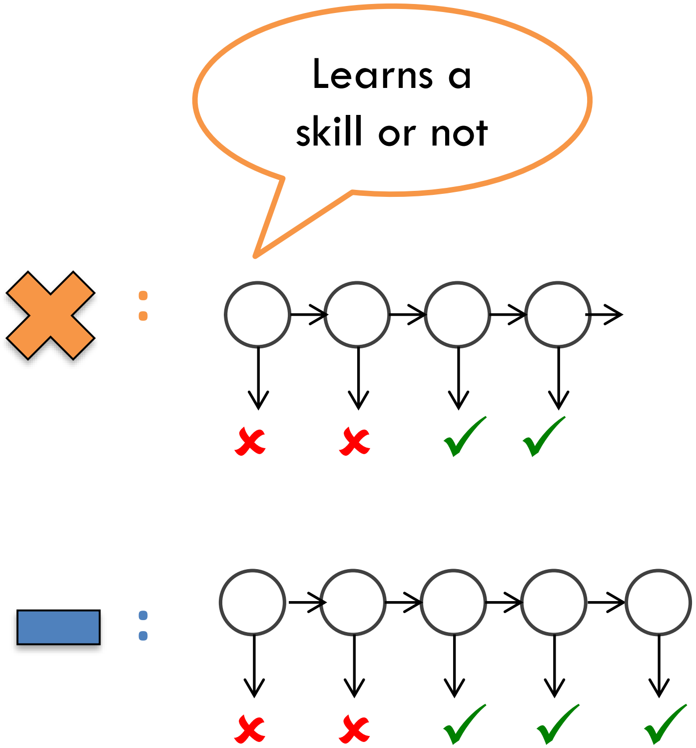
Outline

- Introduction
- The Polygon Evaluation Framework
- Studies and Results
- Conclusions

Motivation

- Data-driven Student Modeling : different “well-fitted” models from the same data
- But, usually only a single model is evaluated
- To illustrate, let’s firstly briefly go through two effective student models: Knowledge Tracing and FAST

Knowledge Tracing



- Knowledge Tracing fits a two-state HMM per skill
- Binary latent variables indicate the knowledge of the student of the skill
- Four parameters:
  1. Initial Knowledge
  2. Learning
  3. Guess
  4. Slip

Transition

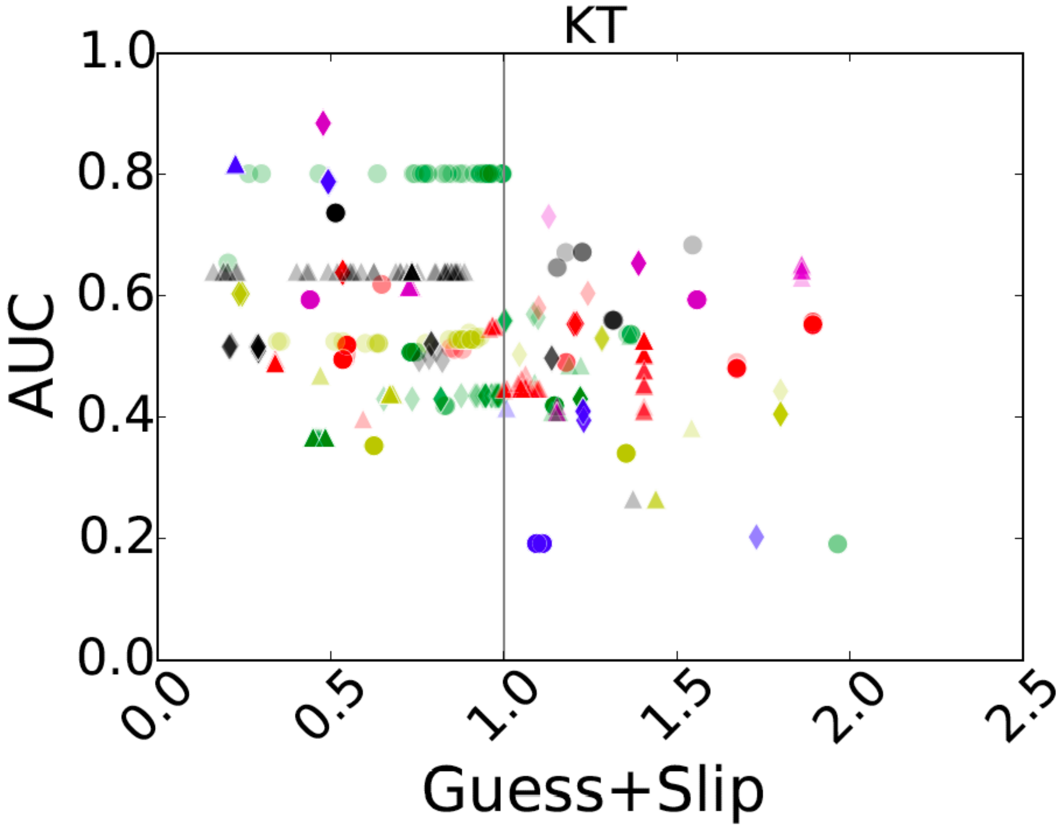
Emission

Feature-Aware Student Knowledge Tracing

- Knowledge Tracing + features
- Features : contextual information
  - Item difficulty
  - Student ability
  - Requested hints?
  - ...
- How do features come in: replacing the binomial distributions by logistic regression distributions.
- Details in our 2014 EDM paper *(General Features in Knowledge Tracing to Model Multiple Subskills, Temporal Item Response Theory, and Expert Knowledge.)*

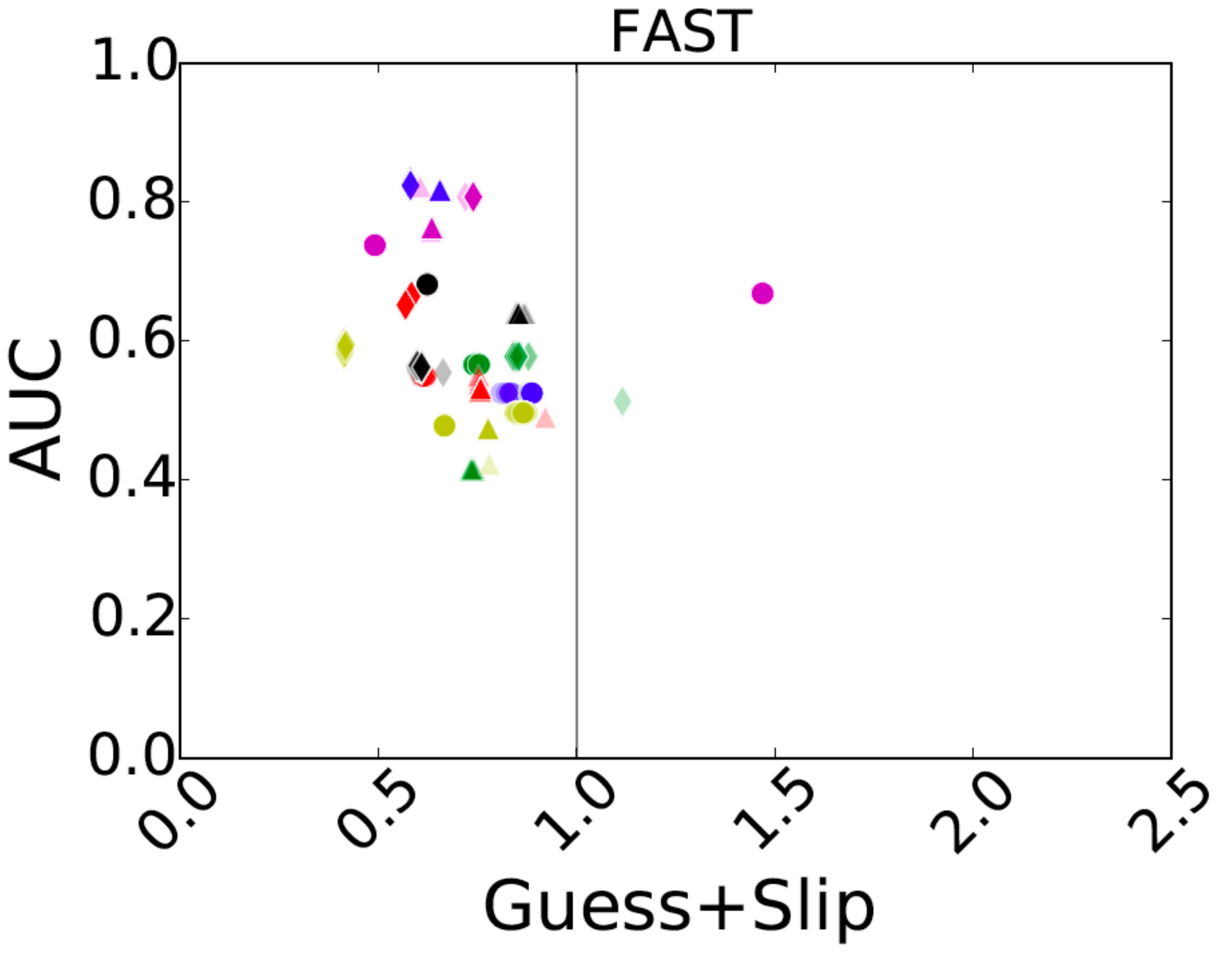
Do we always get a similar model?

- Knowledge Tracing
- A point : best fit model from one run for a skill
- A color-shape : a skill with 100 runs

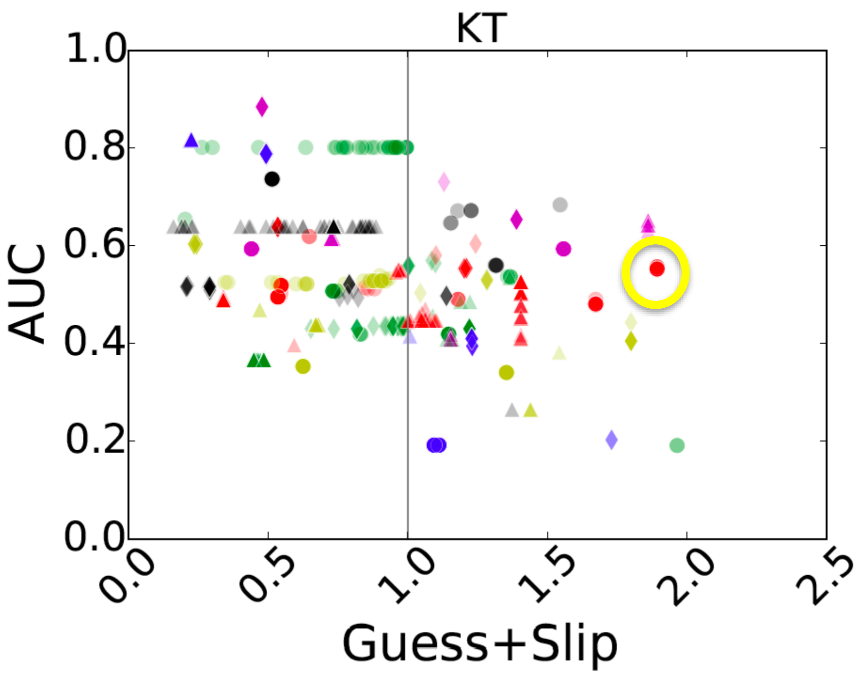


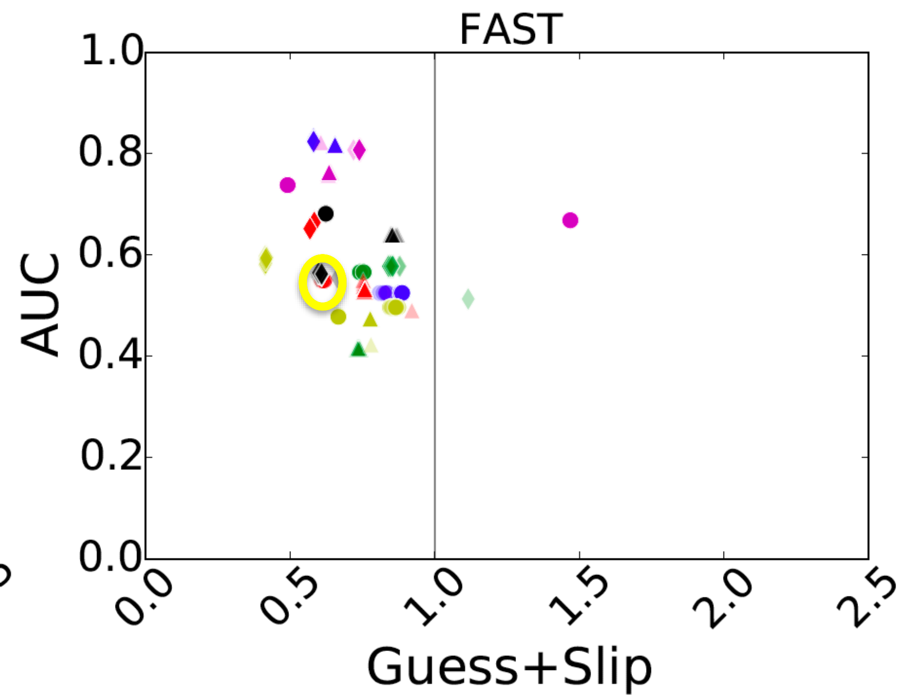
What about a more complex student model?

- Less spreading. Seems to get a single model.



Which modeling approach is better?





- Single model of one skill
- AUC : KT > FAST
- Guess+Slip : Very different! FAST > KT (details later)
- Stability: FAST > KT
- Which modeling approach is better for this skill?

Predictive performance is not enough …

Some literatures pointing out different dimensions can be found for Knowledge Tracing … (consider adding more)

- Beck et al '07 :
  - Identical global optimum predictive models can correspond to different sets of parameter estimates (identifiability problem)
  - Extremely low learning rates are considered implausible.
  - Consider putting his graph?

- Baker et al ‘08 :
  - Sometimes, we get models where a student is more likely to get a correct answer if he/she does not know a skill than if he/she does (model degeneracy problem).
  - Empirical values for detection:
    - The probability that a student knows a skill should be higher than before the student’s first **3** actions.
    - A student should master the skill after **10** correct responses in a row.

- Gong et al ‘10 : do fitted parameters correlate with pre-test scores well?
- Pardos et al ’10 : the optimization algorithm can converge to the local optima yielding different properties of parameters that depend on the initial values (put his graph?)
- De Sande ’13 : Empirical degeneracy can be precisely identified by some theoretical conditions.
- De Sande ’13, Gweon ‘15: presented different (and even contradictory) views of Beck’s identifiability problem.

General problems for latent variable models

- Latent Variable student models: infer student knowledge from performance data
- Finding optimal model parameters is usually a difficult non-convex optimization problem for latent variable models.
  - Many latent variable student models are used to in adaptive tutoring systems to trace student knowledge.
- Moreover, in the context of tutoring systems, even global optimum model parameters may not be interpretable (or plausible).

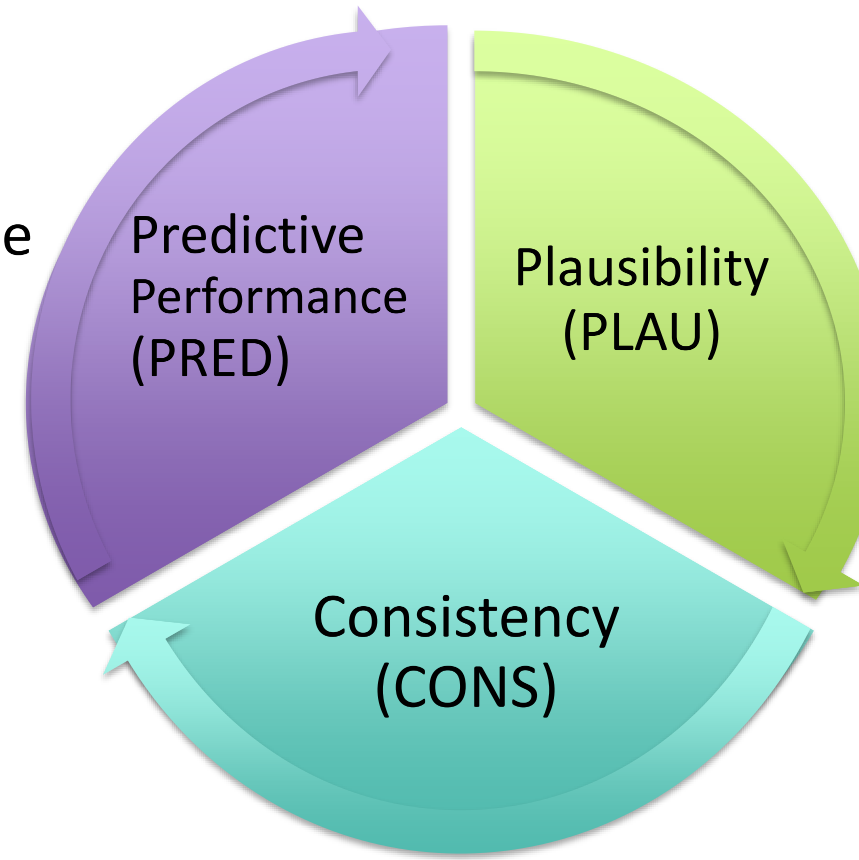
Can we get a unified, generalizable view?

Outline

- Introduction
- The Polygon Evaluation Framework
- Studies and Results
- Conclusions

Polygon: A Multi-faceted Evaluation framework

How well does the model predict?



How interpretable (plausible) are the parameters for tutoring systems?

If we train the model under different settings (later mention), does the model give same (similar) parameters?

Procedurals

1. Define potential metrics to instantiate the framework
2. Run Knowledge Tracing and Feature-Aware Student Knowledge Tracing with 100 random initializations.
3. Metric selection
4. Model examination and comparison in terms of
   - Multiple Random Restarts
   - Single models (details in paper) 

5. Implications for Single Model Selection

Constructing Potential Metrics

- Each metric is computed for one skill (knowledge component, i.e., KC).
  - We then aggregate multiple skills to get the overall picture.
- Each metric can evaluate a single restart model and multiple restart models (except for consistency metrics).
- Each metric ranges from 0 to 1.
- Higher positive value indicating higher quality.

Predictive Performance

- AUC and P-RAUC.
  - Intuition: A good model should predicts well.
  - AUC gives an overall summary of diagnostic accuracy.
    - 0.5: random classier, 1.0: perfect accuracy.
  - Each random restart : AUC^r
  - Across 100 random restarts: P-RAUC

$$\text{P-RAUC} = \frac{1}{R}\sum_{r=1}^{R}\mathrm{AUC}^r \tag{1}$$

Welcome to consider other metrics if you have concerns.

Plausibility

- Guess+Slip<1 (GS) and P-RGS
  - Intuition: A good model should comply with the idea that knowing a skill generally leads to correct performance.
  - De Sande '13 proves a condition guaranteeing Knowledge Tracing not to have empirical degeneration:

$$\mathrm{GS}^r = \underbrace{\mathbb{1}}_{\text{indicator function (0/1)}}(\mathrm{Guess}^r + \mathrm{Slip}^r < 1) \tag{2}$$

  - Across 100 random restarts: P-RGS

$$\text{P-RGS} = \frac{1}{R}\sum_{r=1}^{R} \mathrm{GS}^r \tag{3}$$

Plausibility

- Non-decreasing predicted probability of Learned (NPL) and P-RNPL.
  - Intuition: we take the perspective that a decreasing predicted probability of learned implies practices hurt learning, which is not plausible. *(We are aware of the other perspective where it is interpreted as a decrease in the model's belief.)*
  - This is general to all latent variable models.

D: #datapoints

s: student

t: practice opportunity

O: observed historical practices

$$\mathrm{NPL}^r = \frac{1}{D}\sum_{s=1}^{S}\sum_{t=1}^{T_s-1}\mathbb{1}[\mathrm{P}(\widetilde{L}_{t+1}^{rs}|\mathbf{O}^{rs}) \geq \mathrm{P}(\widetilde{L}_{t}^{rs}|\mathbf{O}^{rs})] \quad (4)$$

$$\text{P-RNPL} = \frac{1}{R}\sum_{r=1}^{R}\mathrm{NPL}^r \quad (5)$$

Consistency

- Intuition: A good model should be more likely to converge to points with higher predictive performance and plausibility, and give more stable predictions and inferences.

  - Consistency of AUC, GS, NPL (C-RAUC, C-RGS, C- RNPL)
    - For example, to compute the consistency of AUC:

$$\text{C-RAUC} = 1 - \underbrace{\sqrt{\frac{1}{R}\sum_{r=1}^{R}(\text{AUC}^r - \overline{\text{AUC}})^2}}_{\text{uncorrected sample standard deviation}} \qquad (6)$$

Consistency

- Consistency of the predicted probability of mastery (C-RPM)
  - We define probability of mastery PM as follows:

$$\mathrm{PM}^r = \overbrace{\frac{1}{S}\sum_{s=1}^{S} \overbrace{\mathbb{1}\{\mathrm{P}(\widetilde{L}_t^{rs}|\mathbf{O}^{rs}) \geq 0.95, \exists t \in [1, T_s]\}}^{\text{whether a student ever reached mastery of a skill}}}^{\text{Percentile of students ever reached mastery of a skill}} \quad (7)$$

  - Across 100 random restarts: C-RPM

$$\text{C-RPM} = 1 - \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\mathrm{PM}^r - \overline{\mathrm{PM}})^2} \quad (8)$$

Consistency

- Cohesion of the parameter vector space (C-RPV)
  - De Sande ’13 used fixed point analysis to show that we need all four parameters to dene the overall behavior of Knowledge Tracing during the prediction phase (when knowledge estimation is updated by prior observations).

$$\text{C-RPV} = 1 - \frac{1}{2R} \sum_{r=1}^{R} \overbrace{||\mathbf{V}^r - \overline{\mathbf{V}}||}^{\text{Euclidean distance}} \qquad (9)$$

$(\mathrm{Init}^r,\ \mathrm{Learn}^r,\ \mathrm{Guess}^r,\ \mathrm{Slip}^r)$

Mean of the vector

Metric Selection

- Allows flexible metrics to instantiate each dimension. Here we present some simple ones.
- A principled way to select metrics:
  - cover all three dimensions
  - having the least overlap.
- We examine the scatterplot and correlation of each pair of the metrics and conduct significance tests.

Outline

- Introduction
- The Polygon Evaluation Framework
- Studies and Results
- Conclusions

Real world datasets

Dataset	#observations	#skills	#students	%correct
Geometry	5,055	18	59	75%
Statics	23,390	17	326	77%
Java	43,696	20	328	67%
Physics	10,063	10	40	62%

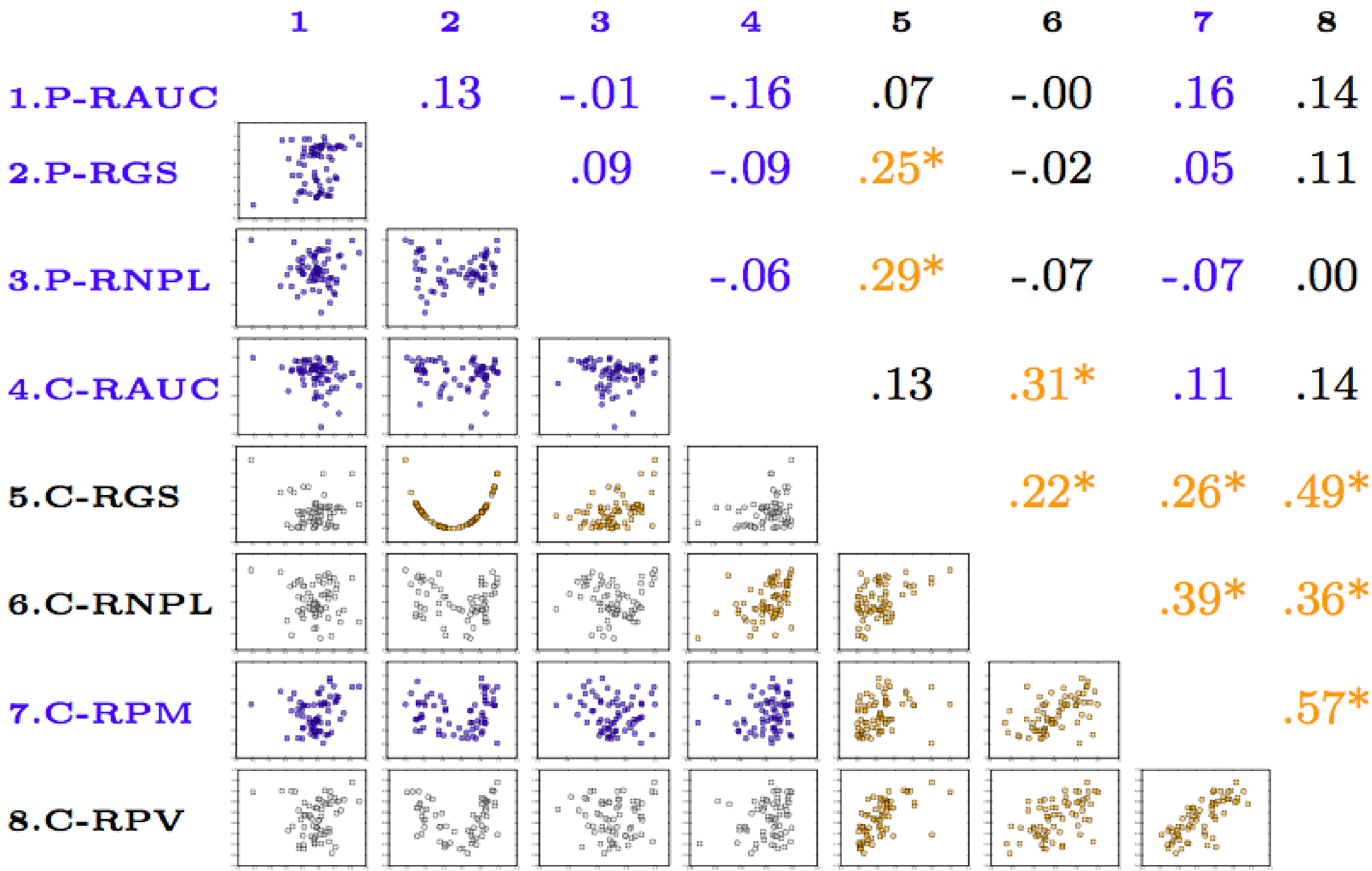
- 65 skills in total
- Geometry: Geometry Cognitive Tutor (Koedinger et al. '10, '14)
- Statics: OLI Engineering Statics (Steif et al. '14, Koedinger et al. '10)
    - Randomly selected 20 skills and removed 3 with #obs< 10
- Java: Java programming tutor QuizJET (Hsiao et al. '10)
- Physics: BBN learning platform (Kumar et al. '15)

Experimental Setup

- Initialize: uniformly at random for 100 times.
  - init, learn, guess, slip: (0, 1)
  - Feature weights: (-10, 10)
- 80% students on train set, remaining on test set.
- Compare standard Knowledge Tracing (KT) and Feature-Aware Knowledge Tracing (FAST) with different features
- FAST:
  - Geometry, Statics, Java: binary item indicator
  - Physics: binary *problem decomposition requested* indicator
  - Features are incorporated into all four parameters (init, learn, guess, slip) in our study.

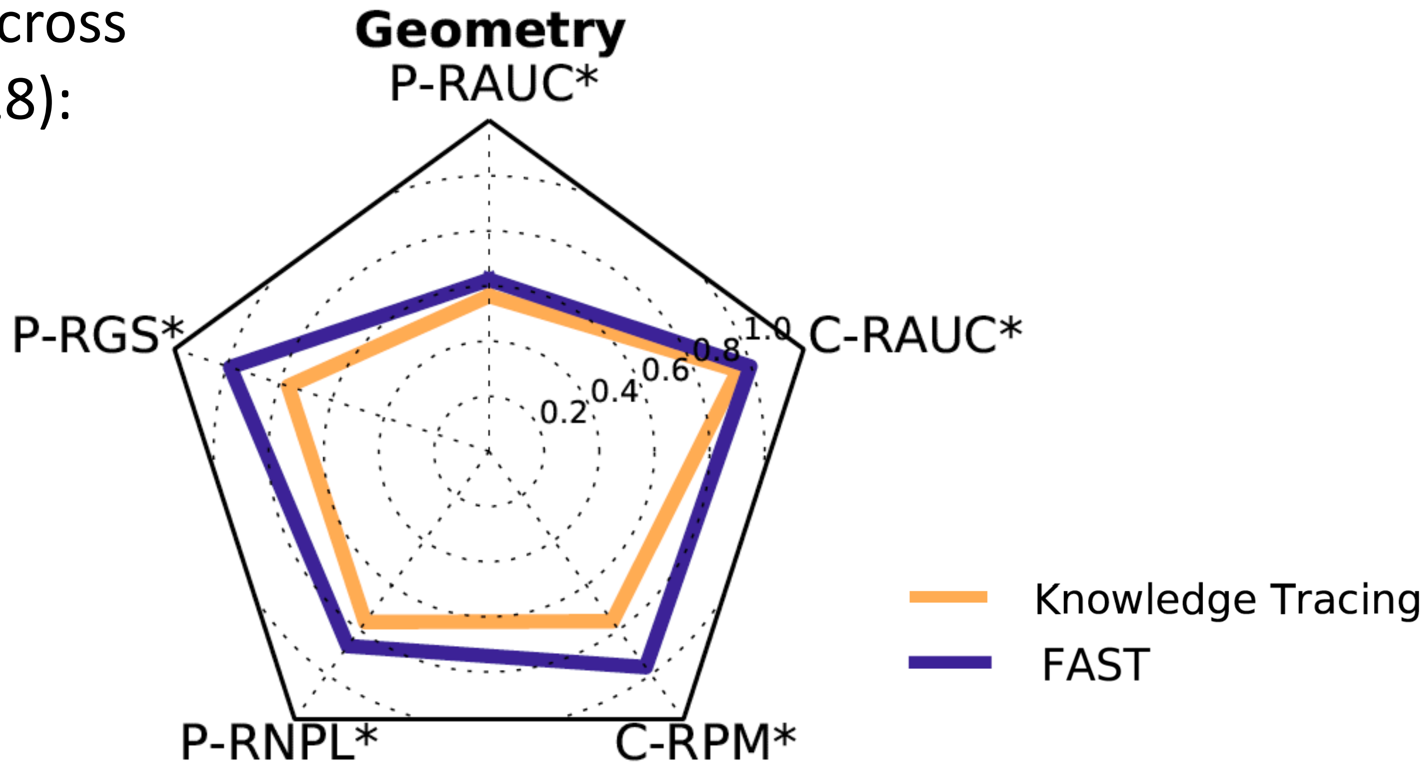
Metric Selection

- Correlation among metrics of all skills (65) from Knowledge Tracing.
- We choose the metrics in blue to instantiate Polygon.



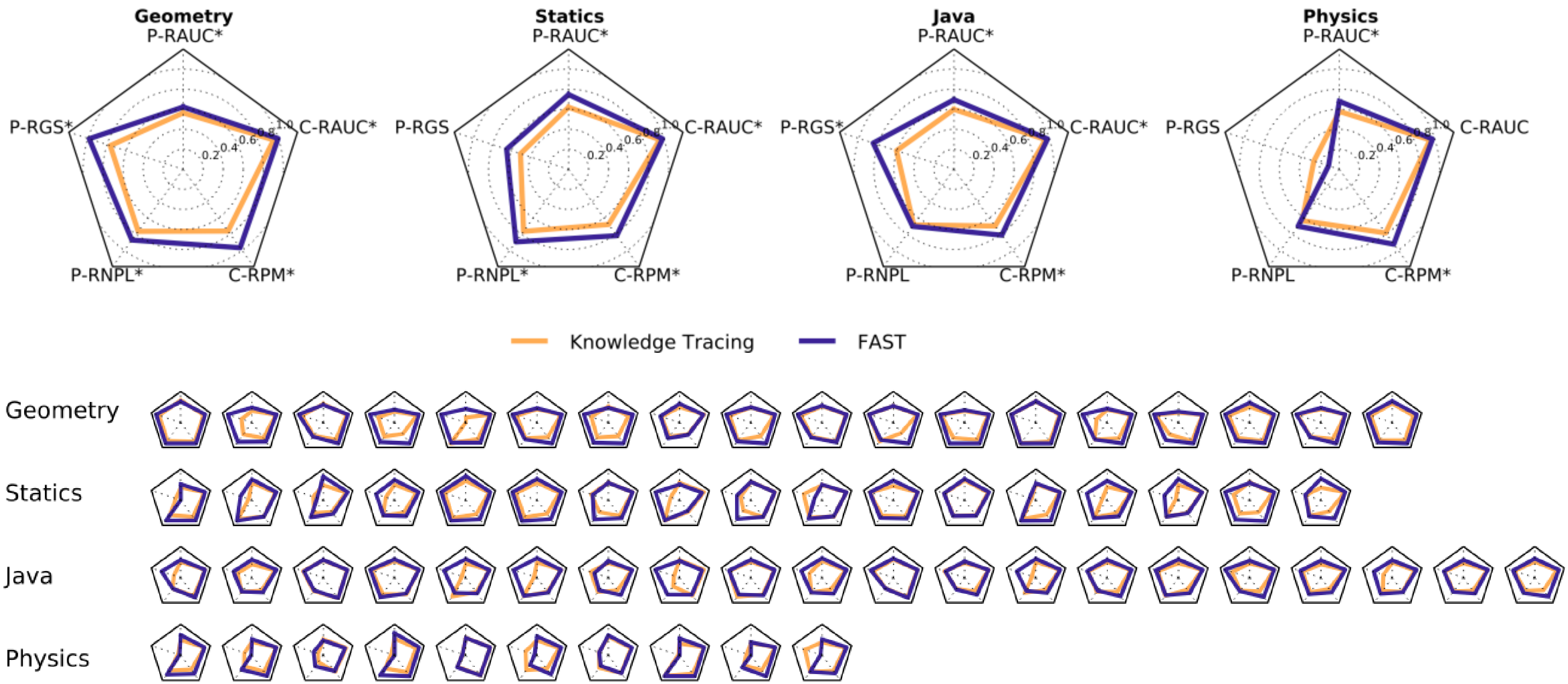
Evaluation on Multiple Random Restarts

- Average across all skills (18):



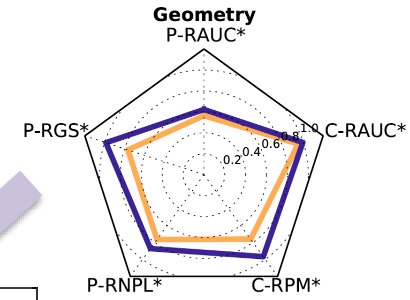
- Individual skills:

Evaluation on Multiple Random Restarts



- FAST's Polygon areas in most cases cover Knowledge Tracing's.
- FAST's plausibility improvement varies across datasets.
    - On Physic dataset, the skill definition may be too coarse-grained and FAST may be more vulnerable to bad skill definitions.

Drill-down Evaluation of Single Models



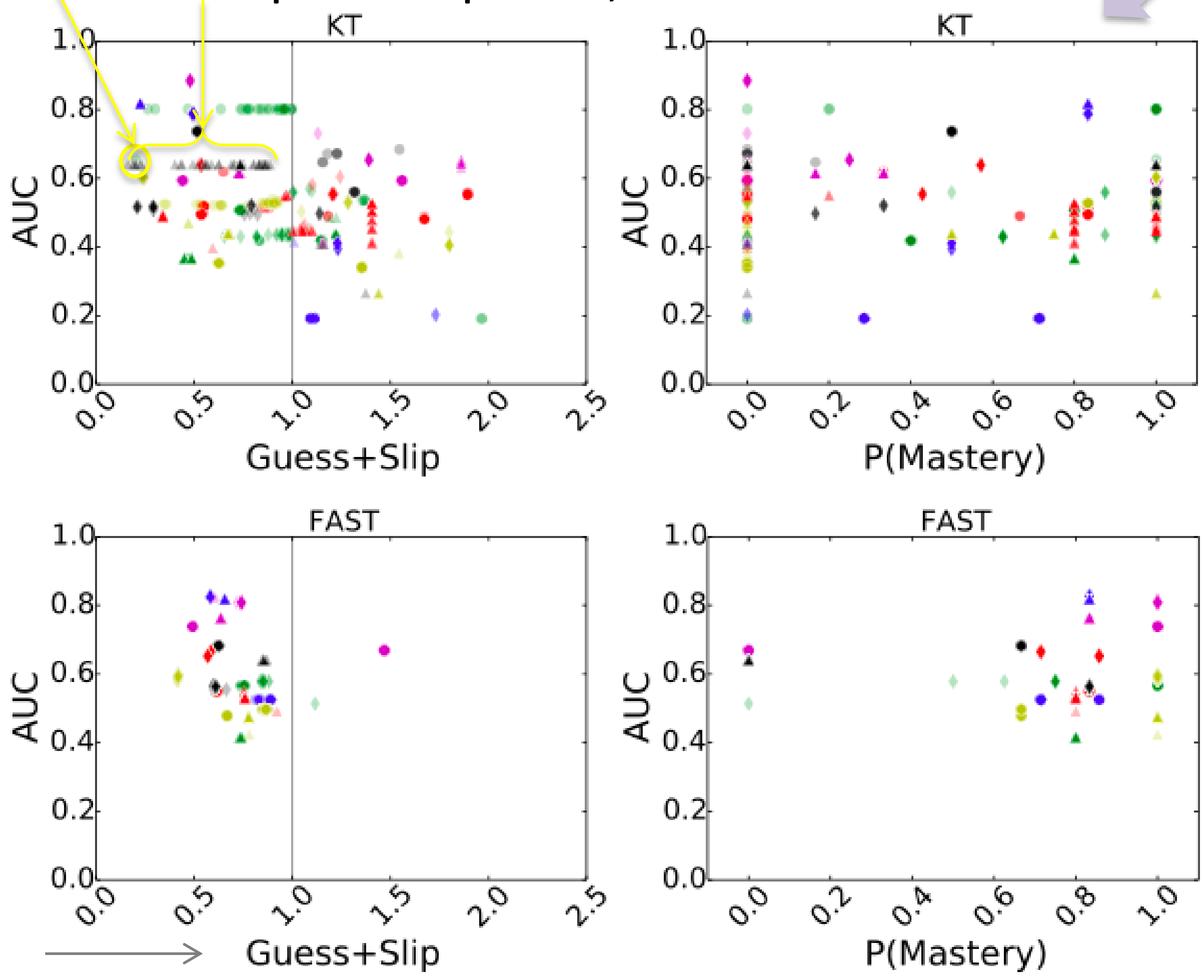
Geometry dataset

Each point: one random restart

Each color-shape: 100 points, 100 restarts

P-RAUC

C-RAUC



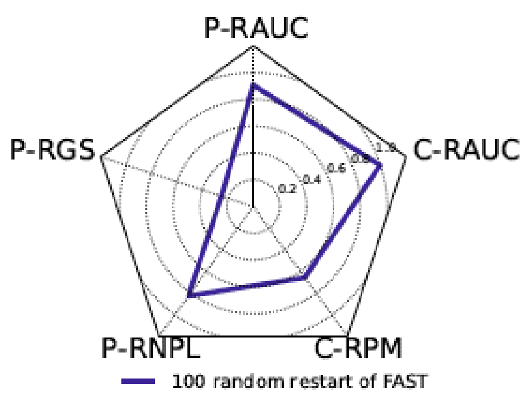
We can also plot **NPL** here

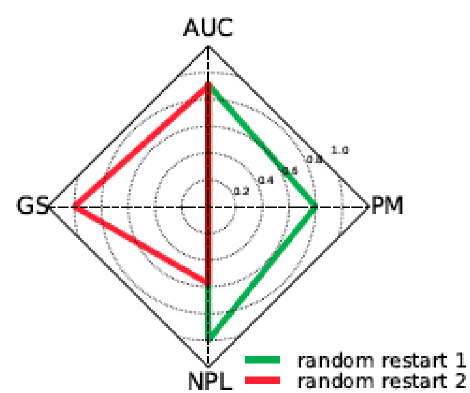
P-RGS (P-RNPL)

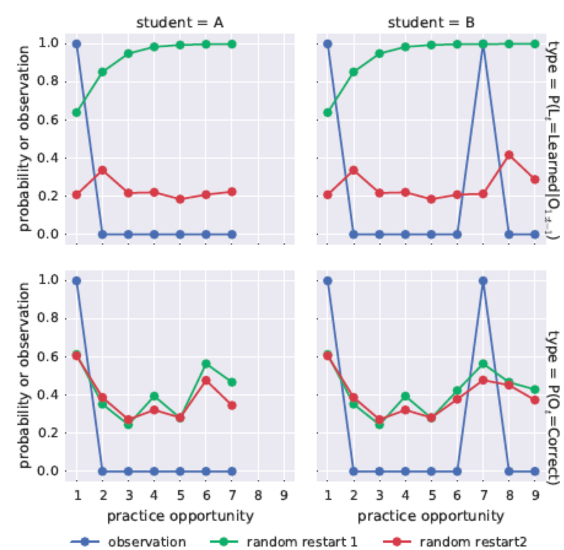
C-RPM

Drill-down Evaluation of Single Models

- FAST comparing with Knowledge Tracing:
  - higher predictive performance
  - more plausible
  - more consistent!
- We also use Polygon framework to effectively identify and analyze skills where FAST is worse than KT on some dimensions. Details in the paper.







How can be choose a single model?

- Choose the random restart with the highest AUC?

	GS		NPL	
	+	−	+	−
AUC	41(0.6)	23(-0.6)	35(0.7)	30(-0.5)

For example, among all 65 skills for Knowledge Tracing, 41 skills have positive correlation between AUC and GS across 100 restarts. The average correlation is 0.6.

- Overall, more than 35% of skills show negative correlations between predictive performance and plausibility with non-trivial magnitude (.5~.6)!

How can be choose a single model?

- Choose the random restart with the highest log likelihood on train set?

	AUC		GS		NPL	
	+	−	+	−	+	−
LL	46(0.5)	19(-0.4)	34(0.5)	30(-0.5)	30(0.4)	35(-0.5)

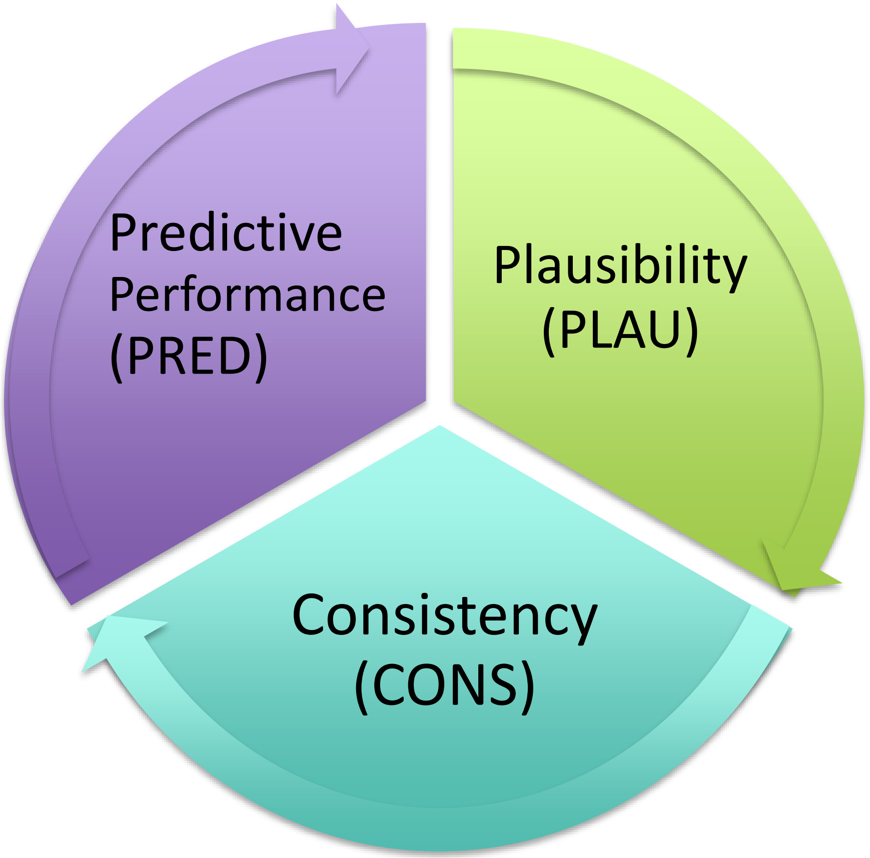
- Similarly, more than 46% of skills show negative correlations between predictive performance and plausibility with non-trivial magnitude (.5)!
- A practical way to select a single model with high quality in all dimensions is still under question.
- Polygon framework provides important insights.

Outline

- Introduction
- The Polygon Evaluation Framework
- Studies and Results
- Conclusions

Contributions

- A unified, general, multifaceted evaluation framework to quantify the quality of student models:



Conclusions

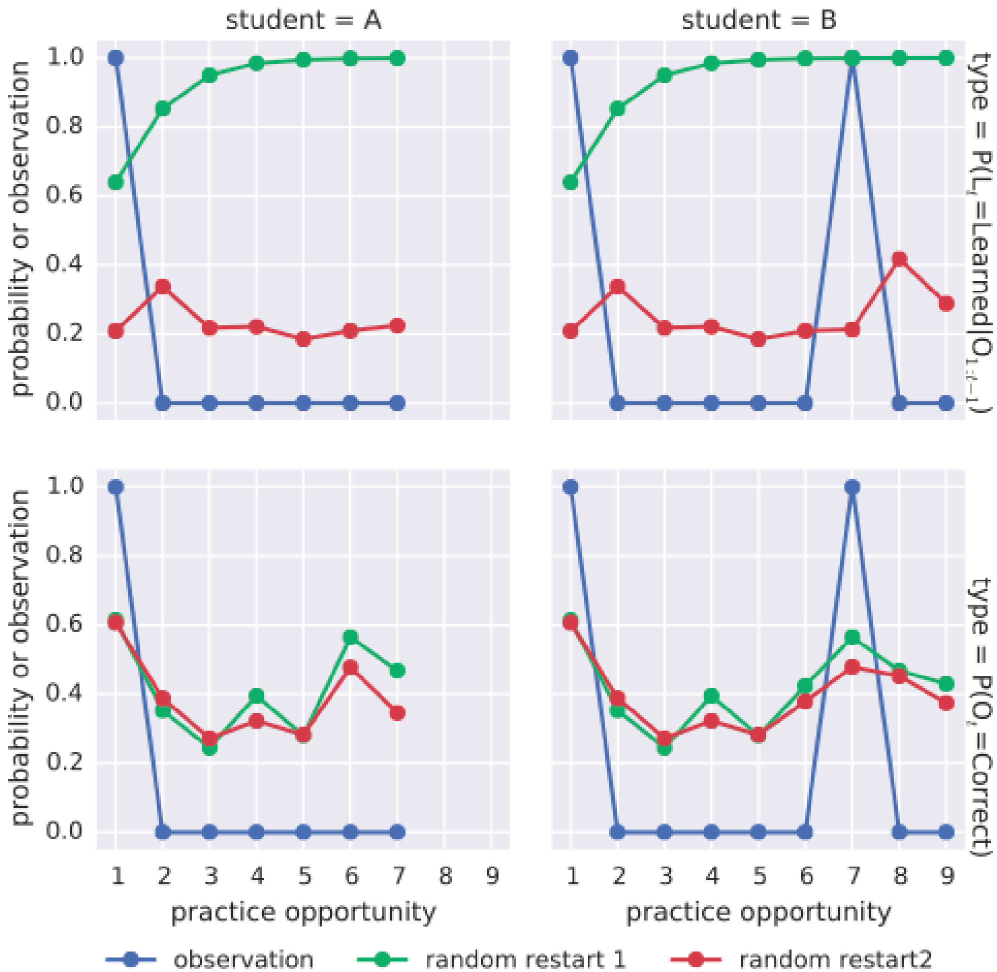
- A recent model FAST with proper features can promise higher predictive performance, plausibility and consistency than Knowledge Tracing.
- One reason can be: Features indirectly constrain the optimization algorithm to search within regions with both high fitness and plausibility.

Conclusions

- Our study is still exploratory and serves as a first step towards more theoretical, deeper understanding of the parameter space of complexed student models.
  - Better metrics? More dimensions?
  - external measurements?
  - decrease or increase the number of random restarts?
  - well-defined vs. ill-defined knowledge components?
  - combine these three dimensions in a single metric?
  - …

Thank you for listening!

Drill-down Evaluation of Single Models



- Extending the identifiability problem: they have very similar predicted correctness, yet present fun- damentally different predicted knowledge levels.
- Also, we observe the empirical degen- eracy of random restart 1: with more incorrect practices, the predicted probability of Learned increases.
- This analy sis showcases the effectiveness of Polygon metrics in identifying hidden problems.