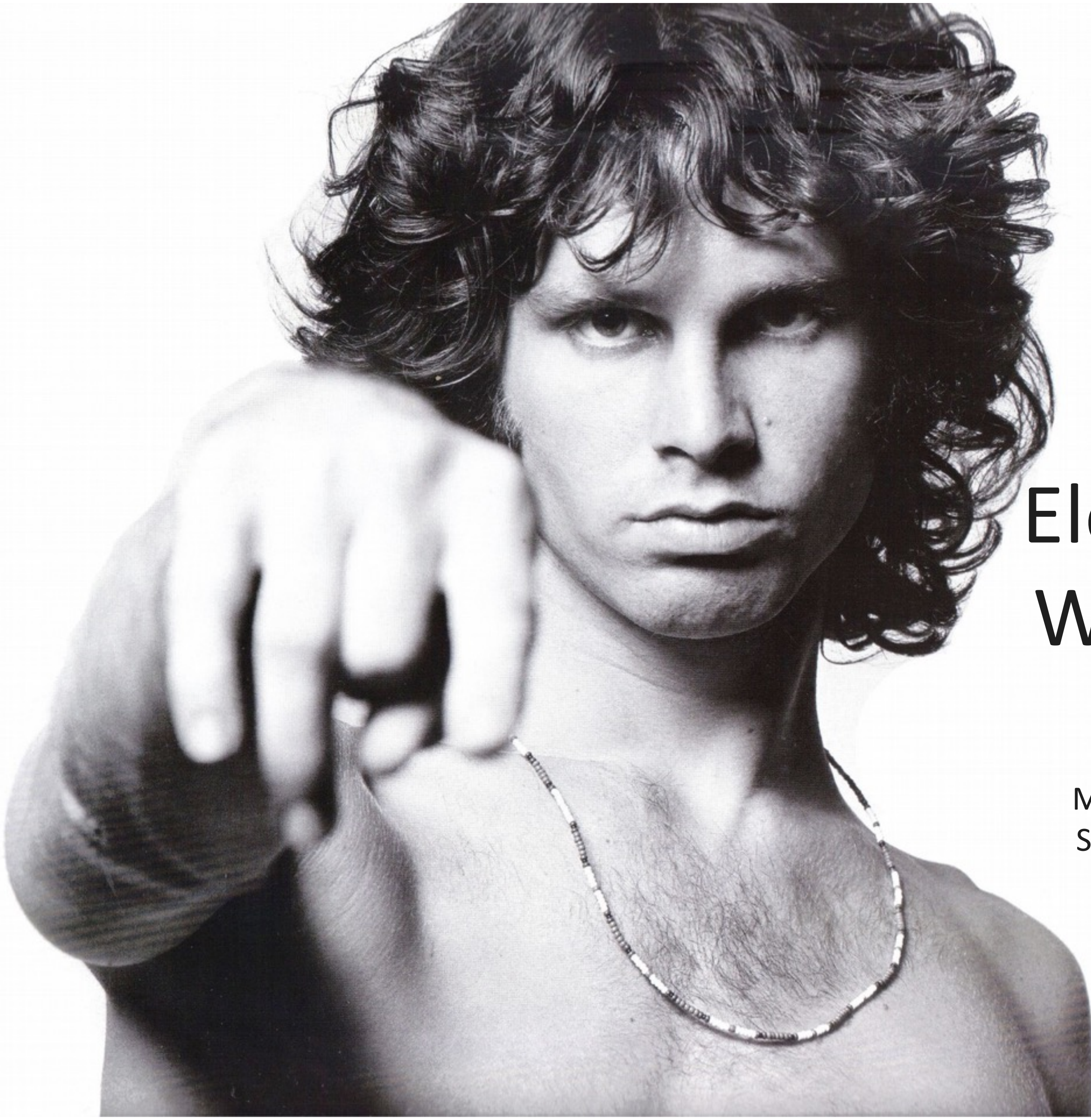
Elo, I Love You, Won't You Tell Me Your K

Michael Yudelson, Sr. Research Scientist, ACTNext by ACT, Inc.

ELO – Electric Light Orchestra



Elo, I Love You, Won’t You Tell Me Your K

Michael Yudelson, Sr. Research Scientist, ACTNext by ACT, Inc.

Elo, I Love You, Won’t You Tell Me Your K

Michael Yudelson, Sr. Research Scientist, ACTNext by ACT, Inc.


ACTNEXT
BY ACT
Home
News
Research and Projects
Events
Team and Partners
About ACTNext
THE ROBOTS ARE COMING
"The Robots Are Coming To School: Now What?"
Posted Nov 28, 2018
Latest Tweets
Kicking off the #CBJtech luncheon w/ opening remarks from @ACT CEO @MartenRoorda about ACT's mission driven transfo...
https://t.co/EST0uOWw0i
— 2 mins ago
May 20-23, get set for #LILI2019, the Learning Impact Leadership Institute hosted by @LearningImpact! Stop by th...
https://t.co/JObqw0JYvy
— 1 hour ago
Featured Team Member
Kristin Stoeffler
Senior Learning Solutions Designer
Presentations
WorldCanvass Podcast Featuring ACTNext Sr. Research Scientist...
Posted Feb 25, 2019

Arpad Elo (Árpád Imre Élő)

Arpad Emmerich Elo the creator of the Elo rating system for two-player games such as chess. Born in Egyházaskesző, Austro-Hungarian Empire, he moved to the United States with his parents in 1913.

Elo rating schema When two opponents with known ratings meet, their ratings are updated given the result of their match (win, loss, draw). Update is greater as the expected outcome is farther from the actual.

ACTNEXT BY ACT

Uses of Elo

Not in Education

- Chess
- Videogames (CounterStrike)
- Competitive Sports: American football, basketball
- Table games: scrabble
- Dating apps: Tinder

In Education

- MathGarden (Hofman et al. 2018)
- Teaching geographic shape recall (Nižnan et al., 2015)
- Biology – explaining primate behaviors (Franz et al., 2015)
- Assessment/Test prep/Learning – ACT Academy + RAD API (Von Davier et al., 2019)

Simple Elo Resembles 1PL IRT (1)

1PL IRT (Rasch model)

- Student ability θ_i
- Item difficulty β_j
- Assume θ_i and β_j are stationary
- Considerable calibration is necessary
- Accurate for high-stakes tests
- Theoretical guarantee

Simple Elo (Student-Item)

- Running estimate of student ability s_i
- Running estimate of question item difficulty b_j
- No such assumption
- Minimal calibration if any
- Ad hoc
- No theoretical guarantee

Simple Elo Resembles 1PL IRT (2)

1PL IRT (Rasch model)

- Student *i* solves item *j*
- Student ability – θ_i (logistic, random)
- Item (question) difficulty – β_j (logistic, fixed)
- Estimate of student success
  - $m_{ij} = \theta_i - \beta_j$
  - $p_{ij} = \frac{1}{1+e^{-m_{ij}}}$ or $p_{ij} = \sigma(m_{ij})$
- Problems to solve
  - Design, calibrate, validate, equate J items with J factors β_j
  - Calibrate 1 shape of distribution of θ_i
  - Given performance, produce θ_i

Simple Elo (Student-Item)

- Student *i* solves item *j*
- Student ability – s_i
- Item (question) difficulty – b_j
- Estimate of student success
  - $m_{ij} = s_i - b_j$
  - $p_{ij} = \frac{1}{1+e^{-m_{ij}}}$ or $p_{ij} = \sigma(m_{ij})$
- Initial value of s_i and b_j = 0
- Update, given observation o_{ij}
  - $s_i = s_i \underline{+ K}(o_{ij} - p_{ij})$
  - $b_j = b_j \underline{- K}(o_{ij} - p_{ij})$
  - K – sensitivity of the update (say, 0.4)
- *Fitting* Elo
  - Grid search for best K
  - BFGS procedure (approximated or analytical gradient)

Not So Simple Elo

- Choice of value classes to track
  - student, item $p_{ij} = \sigma(s_i - b_j)$
  - student, skills $p_{ij} = \sigma(s_i - \sum_k q_{jk}\, b_k)$
  - student, student-skill, skill (hierarchical) $p_{ij} = \sigma(s_i + \sum_k q_{ik}\, s_{ik} - \sum_k q_{jk}\, b_k)$
  - student, student-skill, skill, item
    $$p_{ij} = \sigma(s_i + \sum_k q_{ik}\, s_{ik} - \sum_k q_{jk}\, b_k - b_j)$$
  - student-level values with a $\oplus$, environment-level values with a $\ominus$
- Choice of sensitivity
  - Constant. Global K vs. per-factor K (K_i for students, K_j for items, K_k for skills)
  - Uncertainty $\mathrm{K} = \frac{a}{1+bn}$ – sensitivity depends on number of datapoints *seen*

Proposal

- Consider 2 simple student-item variants of Elo
  - Single sensitivity K (**E1**)
  - Separate sensitivity for students and for items (**E2**)
- Apply Machine Learning paradigm to finding K's
  - Construct a likelihood function (and fit K's using approx.-d gradients)
  - Construct gradients of likelihood (and fit K's using gradients)
- Consider real-life learning data to validate the ML approach
  - 2 datasets, **D1&D2** (small, medium) from Carnegie Mellon's LearnSphere
  - 2 datasets, **D3&D4** (both very large) from KDD Cup 2010
- Compare to traditional learner modeling approach
  - All datasets were collected in Cognitive Tutor where Bayesian Knowledge Tracing was deployed

Gradients of Elo Parameters

$$J = -ln(L_{tot}) = -\sum_{t=1}^{T} (o_t ln(p_t) + (1 - o_t) ln(1 - p_t))$$

$$i = g_i(t), \text{ index of student for row } t$$

$$j = g_j(t), \text{ index of item for row } t$$

$$r_i(l) = r_i(g_i(l)), \text{ time student } i \text{ was seen prior to time } l$$

$$r_j(l) = r_j(g_j(l)), \text{ time item } j \text{ was seenprior to time } l$$

$$c_i = c_i(g_i(l)), \text{ count of times student } i \text{ seen prior to time } l$$

$$c_j = c_j(g_j(l)), \text{ count of times item } j \text{ seen prior to time } l$$

$$m_t = s_i - b_j$$

$$\delta_t = o_t - \sigma(m_t)$$

$$s_i = \begin{cases} 0 & \text{if } c_i = 0 \\ s_i + K \cdot \delta_t & \text{if } c_i > 0 \end{cases}$$

$$b_j = \begin{cases} 0 & \text{if } c_j = 0 \\ b_j - K \cdot \delta_t & \text{if } c_j > 0 \end{cases}$$

$$\frac{\partial J}{\partial K} = -\sum_{t=1}^{T} \delta_t \cdot \sum_{l=1}^{t-1} \left[(c_i > 0) \cdot \delta_{r_i(l)} + (c_j > 0) \cdot \delta_{r_j(l)} \right]$$

Data

	Dataset	N	Students	Items	N/Item
D1	Geometry Area (1996-97)	5,104	59	139	36.72
D2	Geometry Area Study	128,493	123	16,485	7.79
D3	KDD Cup Challenge A	8,918,055	3,310	206,596	43.17
D4	KDD Cup Challenge B	20,012,499	6,043	61,848	323.58

Expectations

- Elo's should not be drastically worse than BKT
  (I knew fit Student-Item Elo beats ***shipped*** BKT on accuracy)
- Single-K Elo would likely loose to Two-K Elo
- Two-K's would center around single-K
- Approximated gradients (in Elo) would be slower than analytical gradients
- Elo parameters from approximated grad.-s would be close to those from analytical grad.-s

Results

Model	Data	Grad.-s	Neg. LL	RMSE	Acc.	Param.(s)	Iter.	Tm., s	Tm./It.
E1	D1	approx.	2639	0.4139	0.7453	0.3583	19	0.022	0.0011
E1	D1	analyt.	2640	0.4140	0.7467	0.3701	60	0.035	0.0006
E2	D1	approx.	2634	0.4137	0.7443	0.2619, 0.4427	25	0.029	0.0012
E2	D1	analyt.	2634	0.4138	0.7437	0.2603, 0.4717	76	0.047	0.0006
BKT	D1	yes	2537	0.4034	0.7663	-	-	0.099	-
E1	D2	approx.	27930	0.2417	0.9299	1.0431	38	0.423	0.0111
E1	D2	analyt.	27957	0.2420	0.9298	0.9381	63	0.687	0.0109
E2	D2	approx.	27269	0.2412	0.9283	0.4128, 1.5169	50	0.738	0.0148
E2	D2	analyt.	27270	0.2411	0.9283	0.4188, 1.5333	137	1.339	0.0098
BKT	D2	yes	29921	0.2500	0.9291	-	-	0.504	-
E1	D3	approx.	3447761	0.3422	0.8538	0.1282	45	22.780	0.5062
E1	D3	analyt.	3450255	0.3422	0.8538	0.0986	49	17.404	0.3552
E2	D3	approx.	3437226	0.3417	0.8539	0.1965, 0.0340	72	40.827	0.5670
E2	D3	analyt.	3440697	0.3421	0.8540	0.1601, 0.0789	152	60.354	0.3971
BKT	D3	yes	3412619	0.3389	0.8572	-	-	46.237	-
E1	D4	approx.	7108867	0.3263	0.8653	0.1212	62	53.871	0.8689
E1	D4	analyt.	7108948	0.3263	0.8653	0.1171	47	38.136	0.8114
E2	D4	approx.	7101767	0.3261	0.8654	0.1697, 0.0734	77	98.708	1.2819
E2	D4	analyt.	7111965	0.3264	0.8652	0.1071, 0.1267	68	65.542	0.9638
BKT	D4	yes	6906909	0.3178	0.8722	-	-	110.052	-

Results

- Elo's should not be drastically worse than BKT
  BKT is better in 0.01x on RMSE, and in 0.01x on Accuracy
- Single-K Elo would likely loose to Two-K Elo
  Totally comparable with differences in 0.001x-0.0001x
- Two-K's would center around single-K
  Confirmed
- Simulated gradients (in Elo) would be slower than analytical gradients
  Well, difference between simulated and analytical gradients is not that straight-forward, but analytical gradients are better on time/iteration
- Elo parameters from simulated grad.-s would be close to those from analytical grad.-s
  Sometimes not, due to differences between simulates/approx. gradients and peculiarities of the ML search procedure (BFGS)

Conclusions

The Good News

- Treating Elo as a ML algorithm
    - Allows for reasoning around it in general ML terms
- A simple two-sensitivity student-item Elo
    - Is extremely simple to implement
    - Is on par with BKT in terms of accuracy
    - Fits faster
    - Remains to be determined if it can replace BKT

The Not So Good News

- More complex Elo versions
  sensitivity K → uncertainty $U_{a,b}$
- Analytical gradients stall fitting
    - Uncertainty = overparameterization?
    - Consider changing ML search method
    - We used BFGS, L-BFGS did not work
- Simulated gradients still work well

Outlook

- Elo rating schema is simple and ***proven to be*** powerful approach
- Describing Elo in terms of likelihood and gradients of parameters given data
  - Useful for operationalizing simple and complex Elo variants
    - 3-variable-class Elo variant is used in ACTNext's RAD API engine
  - Opens up opportunity for Elo-infusion into other models
    - Individualized BKT, where individualization is handled by Elo's s_i and s_{ik} – Elo-infused iBKT
    - Operationalizable AFM & PFA

Thank you!

Assessment, Learning, *and X in between*

Assessment

- Limited timeframe – n<4 hours
- Massed
- Few skill attributes addressed
- Leverage skill attribute mastery covariance

Learning

- Extended timeframe – n>5 hours*
- Spaced
- Many skill attributes addressed
- Focus on higher skill granularity

Ritter, S., Joshi, A., Fancsali, S., & Nixon, T. (2013, July). Predicting standardized test scores from Cognitive Tutor interactions. In *Educational Data Mining 2013*.

1PL IRT

- $E_{ij} = Pr(X_{ij} = 1) = \sigma(m_{ij}) = \sigma(\theta_i - \beta_j) = \frac{1}{1+e^{-(\theta_i - \beta_j)}}$
- High rank (expected)
- *Easy* to fit*
- Precautions
  - Fixed item order (with skill repetition) can boost 1PL IRT rank
  - Descriptive vs. explanatory**

Model	r.Final
MC	8
1PL IRT	4
LLTM	7
AFM	6
PFA	5
BKT	3
Elo S-I	1
Elo S-SK-K	2

* R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification Journal of Machine Learning Research 9(2008), 1871-1874.
** Wilson, M., & De Boeck, P. (2004). Descriptive and explanatory item response models. In *Explanatory item response models* (pp. 43-74). Springer, New York, NY.

LLTM

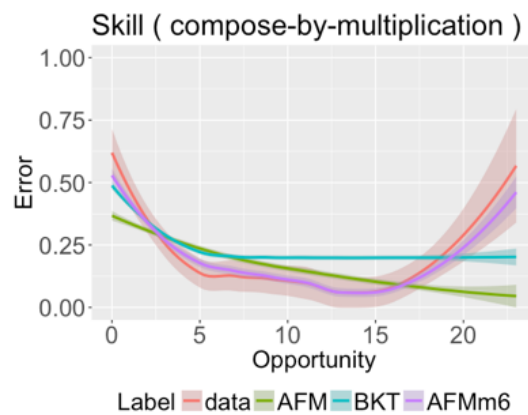
- $m_{ij} = \theta_i + \sum_k q_{jk} \cdot \beta_k$
- In our practice has lower ranks
- *Relatively easy* to fit
- Considerations
  - Sensitive to choice of skill taxonomy*
  - Sensitive to skill indexing (Q-matrix)*
  - Especially as test-prep *moves away* from assessment *closer* to learning
  - Skill tagging in assessment, problem-based learning, and resource recommendation differ (!)

Model	r.Final
MC	8
1PL IRT	4
LLTM	7
AFM	6
PFA	5
BKT	3
Elo S-I	1
Elo S-SK-K	2

* Koedinger, K. R., Stamper, J. C., McLaughlin, E. A., & Nixon, T. (2013, July). Using data-driven discovery of better student models to improve student learning. In *International Conference on Artificial Intelligence in Education* (pp. 421-430). Springer, Berlin, Heidelberg.

AFM & PFA

- AFM: $m_{ij} = \theta_i + \sum_k q_{jk}(\beta_k + \gamma_k t_{ik})$
- PFA: $m_{ij} = \theta_i + \sum_k q_{jk}(\beta_k + \gamma_k s_{ik} + \rho_k f_{ik})$
- *Relatively easy* to fit
- Precautions
  - RE: learning rates *
    - Force γ to be positive
    - Track learning rate magnitudes
  - AFM & PFA have not been operationalized (yet)



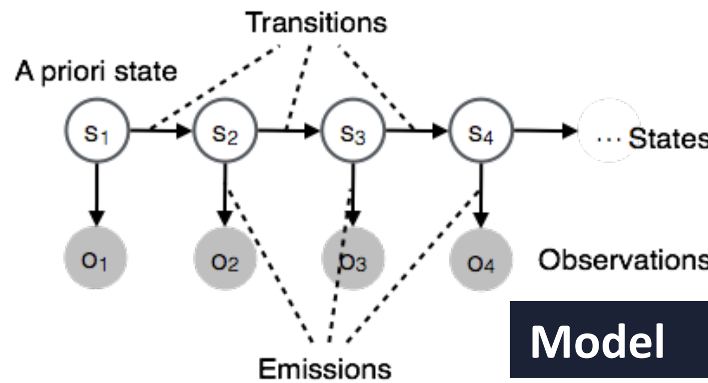
Model	r.Final
MC	8
1PL IRT	4
LLTM	7
AFM	6
PFA	5
BKT	3
Elo S-I	1
Elo S-SK-K	2

* Koedinger, K. R., Yudelson, M. V., & Pavlik Jr, P. I. (2016). Testing theories of transfer using error rate learning curves. *Topics in cognitive science*, *8*(3), 589-609.

ACTNEXT BY ACT

BKT

- $E_{ik} = L_{ik} \cdot (1 - S_k) + (1 - L_{ik}) \cdot G_k$
  $L_{ik} \equiv \sigma(\theta_{ik})$
  $L_{ik}^{t=1} = L_k^0$
  $p(L_{ik}^{t+1}|correct) = \frac{L_{ik}^t \cdot (1-S_k)}{L_{ik}^t \cdot (1-S_k) + (1-L_{ik}^t) \cdot G_k}$
  $p(L_{ik}^{t+1}|wrong) = \frac{L_{ik}^t \cdot S_k}{L_{ik}^t \cdot S_k + (1-L_{ik}^t) \cdot (1-G_k)}$
  $L_{ik}^{t+1} = p(L_{ik}^{t+1}|obs) + (1 - p(L_{ik}^{t+1}|obs)) \cdot T_k$



- Precautions
  - Local optimums in parameter space
  - Label switching
  - If student-skill attempt counts are low, BKT → LLTM without θ_i's; $L_k^0 \rightarrow \beta_k$; S_k, G_k, and T_k assume small random values.

Model	r.Final
MC	8
1PL IRT	4
LLTM	7
AFM	6
PFA	5
BKT	3
Elo S-I	1
Elo S-SK-K	2

Elo Student-Item

- $m_{ij} = s_i - b_j$
  $s_i = 0$, if t=0
  $s_i = s_i + K \cdot (X_{ij} - p_{ij})$, otheriwse
  $b_j = 0$, if t=0
  $b_j = b_j - K \cdot (X_{ij} - p_{ij})$, otheriwse
- Elo r.s. with items most highly ranked
  - ACT Academy dataset(s)
  - Smart Sparrow dataset
  - KDD Cup 2010 challenge B dataset*
- ~~Easy to fit~~; grid search, gradient search**
- Precautions
  - Distributions of s_i and b_j have not been theoretically described and are changing
  - Overfitting is likely but hard to detect***

Model	r.Final
MC	8
1PL IRT	4
LLTM	7
AFM	6
PFA	5
BKT	3
Elo S-I	1
Elo S-SK-K	2

* Stamper, J., Niculescu-Mizil, A., Ritter, S., Gordon, G., Koedinger, K.: Bridge to algebra 2008-2009. challenge data set from KDD cup 2010 educational data mining challenge.
** (under review)
*** Glickman, M. E. (1995). The Glicko system. *Boston University*.

Elo Student-Student/Skill-Skill

- $m_{ij} = s_i + \sum_k q_{jk} \cdot s_{ik} - \sum_k q_{jk} \cdot b_k$

$$s_i^* = s_i + \frac{a_1}{1+c_1 \times n_i}(X_{ij} - E_{ij})$$

$$s_{ik}^* = s_{ik} + \frac{a_2}{1+c_2 \times n_{ik}}(X_{ij} - E_{ij})$$

$$b_k^* = b_k - \frac{a_3}{1 + c_3 \times n_k}(X_{ij} - E_{ij})$$

- Consistently ranked high, second only to item-based Elo r.s.'s
- *Tricky* to fit
- Precautions
  - Parameter space is not smooth (suspicion)

Model	r.Final
MC	8
1PL IRT	4
LLTM	7
AFM	6
PFA	5
BKT	3
Elo S-I	1
Elo S-SK-K	2

Final Comparison

Model	Acc.	RMSE	AUC	r.Acc	r.RMSE	r.AUC	r.Final
MC	0.6352	0.60401	0.5000	8	8	8	8
1PL IRT	0.6630	0.46254	0.6734	4	4	4	4
LLTM	0.6577	0.46862	0.6385	5	7	7	7
AFM	0.6555	0.46856	0.6424	6	6	6	6
PFA	0.6546	0.46849	0.6431	7	5	5	5
BKT	0.6747	0.45871	0.6751	3	3	3	3
Elo S-I	0.7189	0.43126	0.7642	1	1	1	1
Elo S-SK-K	0.7159	0.43525	0.7523	2	2	2	2

Averages across 10=5x2 fold runs

Compare to Learning Data

Model	Acc.	RMSE	AUC	r.Final	Acc.**	RMSE	AUC	r.Final
MC	0.6352	0.60401	0.5000	8	0.8617	0.3719	0.5000	6
1PL IRT	0.6630	0.46254	0.6734	4	0.8644	0.3270	0.7386	4
LLTM	0.6577	0.46862	0.6385	7				
AFM	0.6555	0.46856	0.6424	6	0.8631	0.3677	0.6682	5
PFA	0.6546	0.46849	0.6431	5				
BKT	0.6747	0.45871	0.6751	3	0.8722	0.3177	0.7313	2
Elo S-I	0.7189	0.43126	0.7642	1	0.8653	0.3263	0.7381	3
Elo S-SK-K	0.7159	0.43525	0.7523	2	0.8711	0.3120	0.8011	1
	ACT Academy				KDD Cup 2010 Challenge Set B*			

* KDD Cup 2010, Challenge set B; 20,012,499 transactions, 6,043 students, 61,848 problems

** 0.8617 – avg. succ. rate is deceptive, if all step attempts are considered it drops to 0.6435