Adaptive Testing using a General Diagnostic Model

Jill[1]-Jênn[2] Vie[3] Fabrice Popineau[1]
Yolaine Bourda[1] Éric Bruillard[2]

[1] CentraleSupélec, Gif-sur-Yvette
[2] ENS Cachan/Paris-Saclay
[3] Université Paris-Saclay

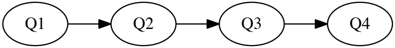
Context

We consider dichotomous data of learners over questions or tasks.

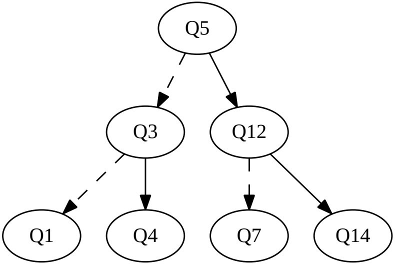
	Questions							
	1	2	3	4	5	6	7	8
Alice	0	1	1	1	0	0	0	1
Bob	1	0	1	1	0	0	0	1
Charles	1	0	1	0	0	0	0	0
Daisy	1	0	0	1	1	1	1	1
Everett	1	0	0	0	1	0	0	1
Filipe	0	1	0	1	1	1	1	1
Gwen	0	0	0	1	0	0	1	1
Henry	0	0	0	0	1	0	0	1
Ian	1	1	1	1	0	1	1	0
Jill	0	1	1	1	0	0	1	0
Ken	1	1	1	0	1	1	0	1

- Tests are too long, students are overtested
- Asking all questions to every learner $\rightarrow$ boredom

How to personalize this process?



Non-Adaptive Test



Adaptive Test

Computerized Adaptive Testing (CAT)

Choose the next question based on previous answers.
$\Rightarrow$ Reduce test length while providing an accurate measurement.

While some termination criterion is not satisfied
  Ask the "best" next question

Psychometry, item response theory (summative)

- Answers can be explained by continuous hidden variables
- What parameters can we measure to predict performance?
- Infer them directly from student data

Cognitive models (formative)

- Answers can be explained by the mastery or non-mastery of some knowledge components (KC)
- Expert maps KCs and items
- Infer the KCs mastered $\Rightarrow$ predict performance

Applications of test-size reduction

- How to ask k questions only, that have predictive power over the rest of the test?
- i.e., k questions that summarize the question set.

Low-stake self-assessment

- Learners get feedback: the KCs that are mastered
- Filter the KCs before assessment
- Practice testing benefits learning (Dunlosky, 2013)

Adaptive pretest at the beginning of a MOOC

- *You seem to lack KCs 1 and 3 that are prerequisites of this course.*
- Personalize course content accordingly
- Recommend relevant resources

Our questions

- How to use a test history data to provide shorter assessments?
- What adaptive testing models exist?
- How to compare them on the same real data?

Outline

- Summative CATs (1983) and formative CATs (2008)
- Comparison framework
- Our new model: GenMA

Summative CATs for standardized tests (GMAT, GRE)

Rasch model for 20 questions

	Q1	Q2	Q3	$\cdots$	Q19	Q20
Difficulty	–0.45	–0.40	–0.35	$\cdots$	0.45	0.50

Question 10 is asked. Incorrect. $\Rightarrow$ Ability estimate $= -0.401$
Question 2 is asked. Correct! $\Rightarrow$ Ability estimate $= -0.066$
Question 9 is asked. Correct! $\Rightarrow$ Ability estimate $= 0.224$
Question 14 is asked. Correct! $\Rightarrow$ Ability estimate $= 0.478$

Feedback and inference

Your ability estimate is 0.478.

- Q1–7 can be solved with proba 0.7
- Q8–15 can be solved with proba 0.6
- Q16–20 can be solved with proba 0.5

Formative CATs for cognitive diagnosis

DINA model for 4 tasks, 4 KCs + slip / guess

		Knowledge components			
		form	**mail**	**copy**	**url**
T1	Sending a mail	**form**	**mail**		
T2	Filling a form	**form**			
T3	Sharing a link			**copy**	**url**
T4	Entering a URL	**form**			**url**

Task 1 is assigned. Correct!

⇒ **form** and **mail** may be mastered. No need to assign Task 2.

Task 4 is asked. Incorrect.

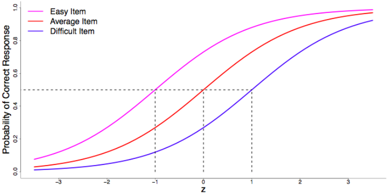
⇒ **url** may not be mastered. No need to use Task 3.

Feedback and inference

- *You master **form** and **mail** but not **url**.*
- *You should read my book on the subject. It's only $200.*

Comparison between summative and formative models

Rasch model



- Difficulty of questions
- Ability of learners
- Learners can be ranked
- No need of domain knowledge

Cognitive diagnosis

	C_1	C_2	C_3
Q_1	1	0	0
Q_2	0	1	1
Q_3	1	1	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$

- KCs required for each question
- Mastery or non-mastery of every KC for each learner
- Learners get feedback
- No need of prior data

GenMA: combining MIRT and a q-matrix

Rasch model

- Perf. depends on difference between learner ability and question difficulty
- Same as Elo ratings

Pr. of success i over j

$$\Phi(\theta_i - d_j)$$

Multidimensional Item Response Theory

- Depends on correlation between ability and question parameters
- Hard to converge

$$\Phi(\vec{\theta_i} \cdot \vec{d_j}) = \Phi\left(\sum_{k=1}^{d} \theta_{ik} d_{jk}\right)$$

$(\theta_{ik})_k$: ability of learner i
$(d_{jk})_k$: difficulty of question j

GenMA

- Depends on correlation between ability and question parameters, but only for non-zero q-matrix entries
- Easy to converge

$$\Phi\left(\sum_{k=1}^{d} \theta_{ik} q_{jk} d_{jk} + \delta_j\right)$$

$(q_{jk})_k$: q-matrix entry
δ_j: bias of question j

Experimental protocol

		Questions							
		1	2	3	4	5	6	7	8
Train	Alice	0	1	1	1	0	0	0	1
	Bob	1	0	1	1	0	0	0	1
	Charles	1	0	1	0	0	0	0	0
	Daisy	1	0	0	1	1	1	1	1
	Everett	1	0	0	0	1	0	0	1
	Filipe	0	1	0	1	1	1	1	1
	Gwen	0	0	0	1	0	0	1	1
Test	Henry	0	0	0	0	1	0	0	1
	Ian	1	1	1	1	0	1	1	0
	Jill	0	1	1	1	0	0	1	0
	Ken	1	1	1	0	1	1	0	1

- Train student set 80%
- Test student set 20%
- Validation question set 25%

Performance evaluation

.6	.1	.6	.7	.9	.1	.5	.5	.3	.7	.9	.4	.1	.6	.6	.7	.3	.7	.6	.3
						T													

2 correct predictions over 5 $\rightarrow$

.8	.4	.8	.6	.4
F	F	T	F	T

.6	.7	.6	.7	.9	.2	.6	.7	.4	.8	.9	.5	.6	.9	.9	.8	.4	.8	.6	.4
						T					F								

3 correct predictions over 5 $\rightarrow$

.6	.4	.8	.4	.4
F	F	T	F	T

Actually, we use log loss:

$$logloss(y^*, y) = \frac{1}{n} \sum_{k=1}^{n} \log(1 - |y_k^* - y_k|).$$

GenMA

Feedback

- The estimated ability $\vec{\theta_i} = (\theta_{i1}, \ldots, \theta_{iK})$
- Proficiency over several KCs

Inference

- Compute the probability of success over the remaining questions

Example

- After 4 questions have been asked
- Predicted performance: $[.62, .12, .42, .13, .12]$
- True performance: $[T, F, T, F, F]$
- Computed logloss (error) is 0.350.

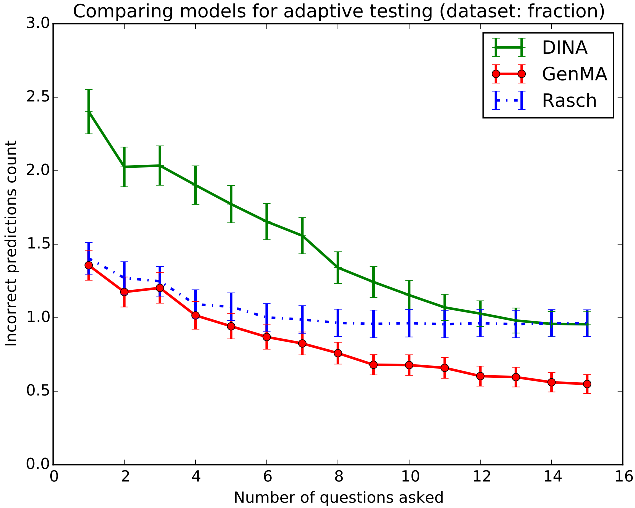
Real dataset: Fraction subtraction (DeCarlo, 2010)

- 536 middle-school students
- 20 questions of fraction subtraction
- 8 KCs

Description of the KCs

- convert a whole number to a fraction
- simplify before subtracting
- find a common denominator
- ...

Results



4 questions over 15 are enough to get a mean accuracy of $4/5$.

Summing up

Rasch model

- Really simple, competitive with other models
- But unidimensional, needs prior data, not formative

DINA model

- Formative, can work without prior data
- Needs a q-matrix

GenMA

- Multidimensional
- Formative because dimensions match KCs
- Needs a q-matrix and prior data
- Faster convergence than MIRT

Further work

Considering graphs of prerequisites over KCs

Attribute Hierarchy Model, Knowledge Space Theory.

Adapting the process according to a group of answers

Multistage Testing.

Doing a pretest with a group of questions, then a CAT

So that first estimate has less bias.

Considering other interfaces for assessment

Evidence-Centered Design, Stealth Assessment (Shute, 2011)

Thank you for your attention!

`github.com/jilljenn`

`jjv@lri.fr`

Do you have any questions?