An empirical approach to on-line learning in SIETTE

Ricardo Conejo,
Eva Millán,
José-Luis Pérez-de-la-Cruz,
Mónica Trella

Department of Computer Science
University of Malaga

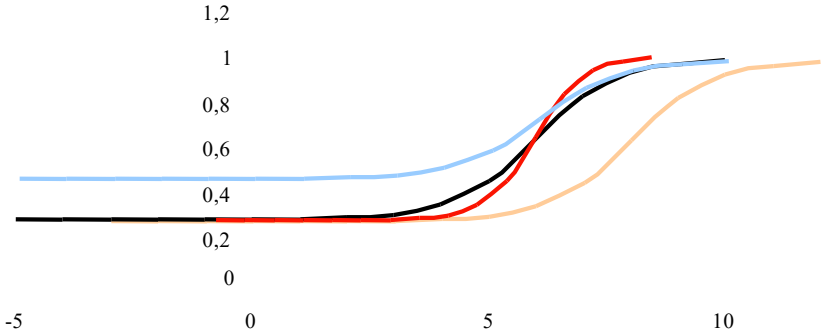
***Item Response Theory* (IRT) & *Computer Adaptive Tests* (CAT)**

• The student's knowledge estimator is a random variable k that takes real values in the interval $(-\infty,+\infty)$.

• Each question or item is assigned a function (*Item Characteristic Curve*, ICC(k)) that represents the probability of answering to it correctly given the student's knowledge level. ICC(k) is assumed to be the normal or the logistic distribution function:

$$ICC(k) = P(C_i^+ \mid k) = c_i + (1 - c_i)\frac{1}{1 + e^{-1.7 a_i (k - b_i)}}$$

where a_i = *discriminnat factor*, b_i = *difficulty factor*, and c_i = *guessing factor*



$a_i = 1.2;$
$b_i = 6;$
$c_i = 0.3$

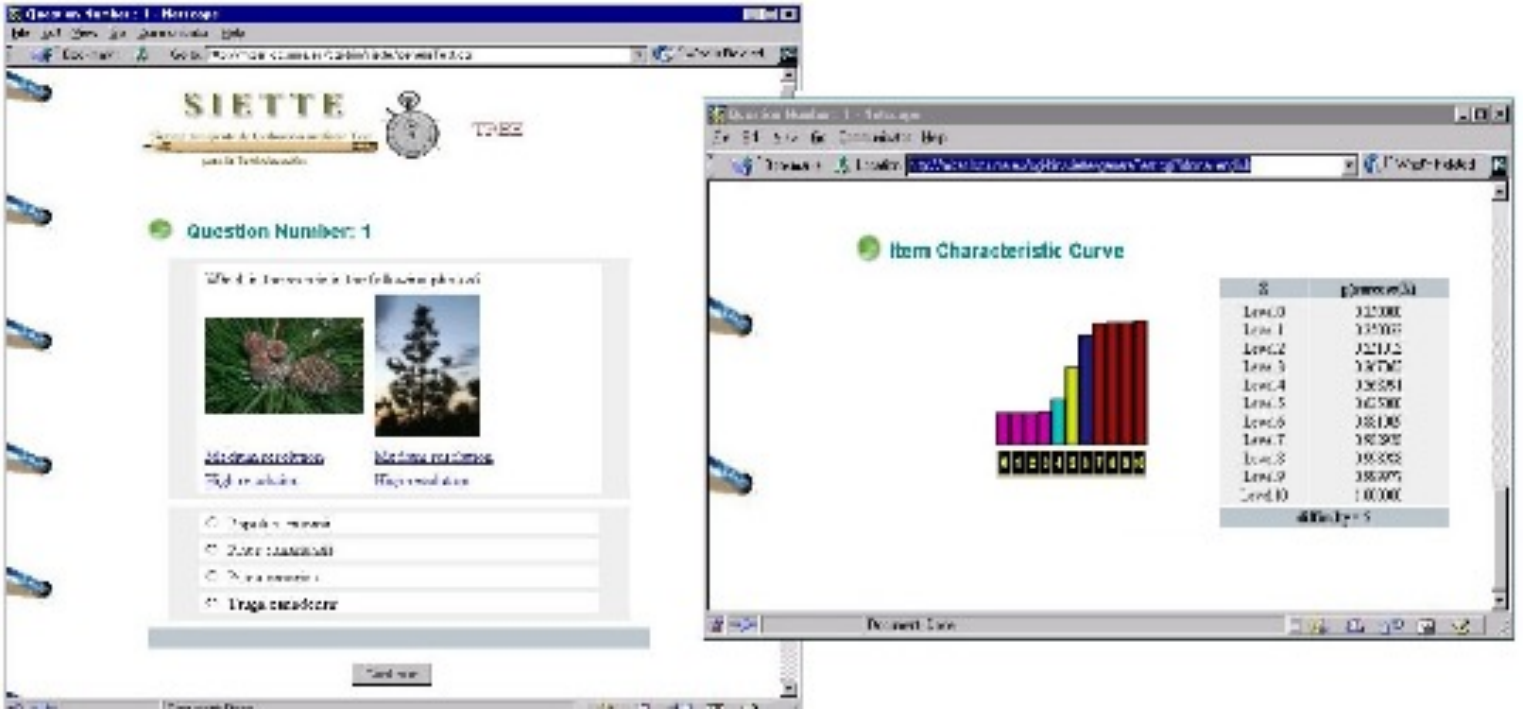
• Evaluation is done by a step-by-step application of the Bayes' rule given the a-priori student's knowledge distribution and the ICC(k) of the question posed.

• CAT = IRT + questions selection criterion + finalisation criterion

SIETTE

SIETTE is a discrete implementation of IRT/CAT for the WWW that can be used as an evaluation module of a *Web based ITS*.

- Teachers can include new questions to existing tests through the WWW interface.
- ICCs are approximated by logistic functions according to parameters given by teachers.
  - Real ICCs are not always logistic functions.
  - Difficulty parameters are not always correctly estimated by teachers.
  - Disciminnant parameters are meaningless for teachers.



Student, item and test simulation

• The student's knowledge estimator is a random variable k that takes integer values between $[0 .. K_{max}]$

• An ICC is given by $K = K_{max} + 1$ values, corresponding to the conditional probabilities of correctly answering the question given that the student belongs to each of the K classes

• The simulation begins with the random generation of a population of N students whose knowledge k is considered uniformly distributed.

• The simulator uses a set of n questions. Their ICC_R of these are generated by assigning values to the parameters a_i, b_i, and c_i in the logistic distribution function.

• A test is simulated for each student, each test contains n questions and the decision of correct/wrong answer is taken according to the student real knowledge k and the ICC_R (k) of the question.

•The confidence factor ρ is the probability of classifying a student correctly after a test.

Simulating a correct calibrated item pool

Accuracy of IRT approximation

ICC_R (a_i = 1.2, b_i uniformly distributed between [1 .. K_{max-1}], c_i = 0)

	Confidence factor ρ = 0.75		*Confidence factor ρ = 0.90*		*Confidence factor ρ = 0.99*	
Number of classes K	*% of correctly classified students*	Average number of questions posed T	*% of correctly classified students*	Average number of questions posed T	*% of correctly classified students*	Average number of questions posed T
3	84.05	2.00	95.82	3.58	99.46	5.65
5	81.61	6.23	92.76	10.38	99.37	19.27
7	80.96	11.11	92.85	18.16	99.38	33.12
9	80.86	16.15	92.93	26.39	99.42	47.27
11	80.52	21.19	92.92	34.54	99.26	60.85

Analysis:

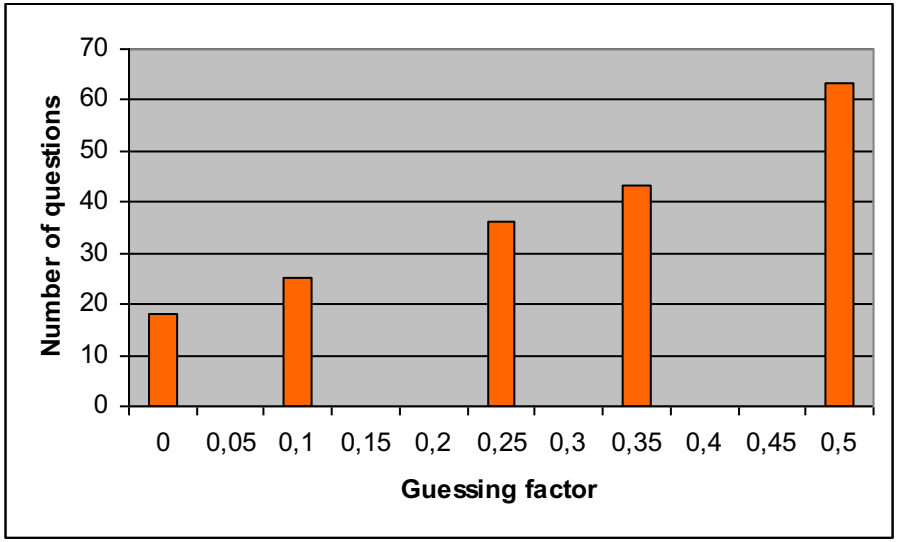
Even with a correct callibrated item pool it is not possible to reach 100% of correctly calibrated students

Guessing factor influence

$\rho = 0.90$; $K=7$

ICC_R ($a_i = 1.2$, b_i uniformly distributed between $[1 .. K_{max-1}]$, $c_i = ...$)

Guessing factor c	*% of correctly classified students*	*Average number of questions posed T*
0.00	92.85	18.16
0.10	92.37	25.34
0.25	92.11	36.05
0.33	91.73	43.37
0.50	91.49	63.37



Analysis :

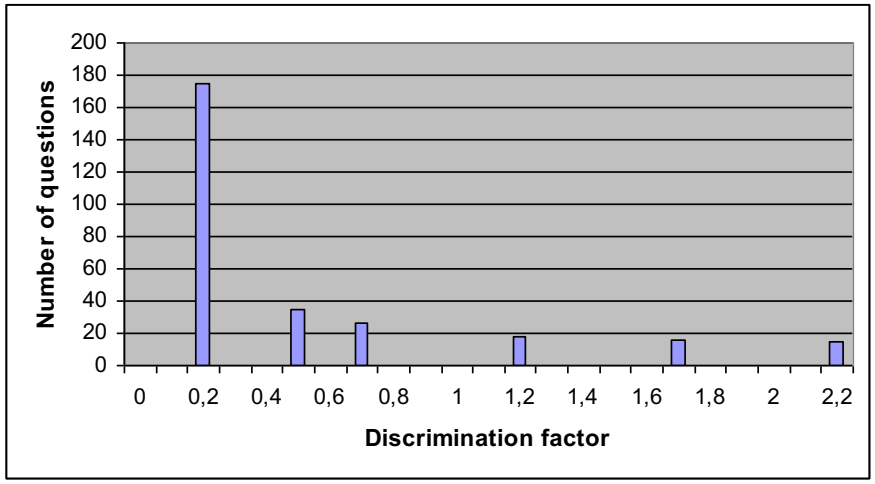
There is no significant influence of the guessing factor in the percentage of correctly clasiffied student, but the number of questions needed increases substantially.

Discrimination factor influence

$\rho = 0.90$; $K=7$

ICC_R (a_i = ... , b_i uniformly distributed between [1 .. K_{max-1}], $c_i = 0$)

Discrimination factor a	*% of correctly classified students*	*Average number of questions posed*
0.20	90.4	174.9
0.50	91.5	35.2
0.70	91.9	26.3
1.20	92.8	18.1
1.70	93.8	15.3
2.20	95.4	14.8



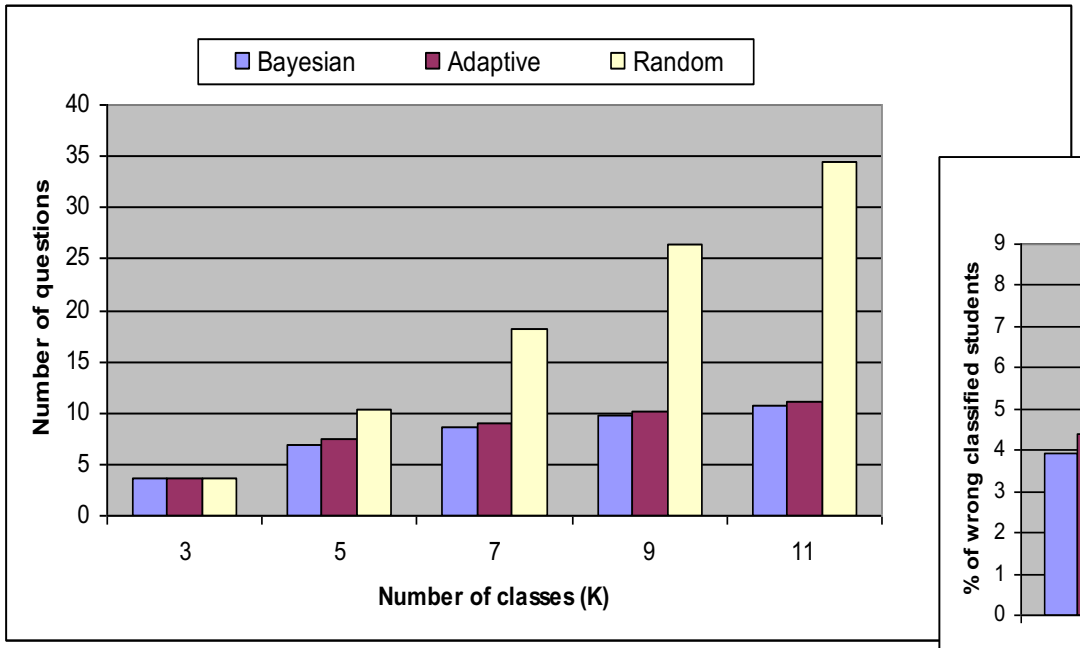
Analysis :

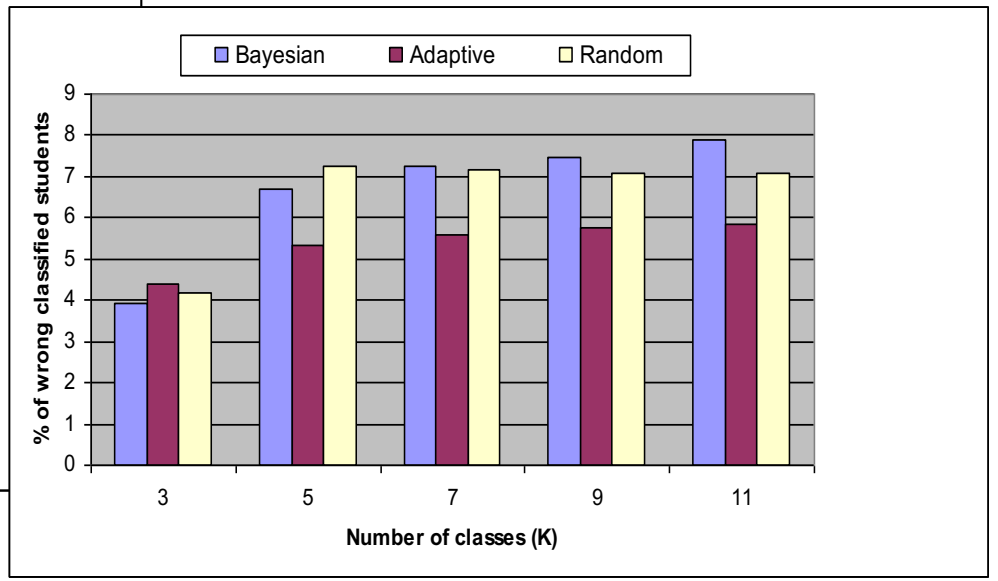
The discrimination factor does not have a great influence in the number of questions if it is bigger than certain threshold.

Accuracy of the CAT

$\rho = 0.90$; K= ...

ICC_R ($a_i = 1.2$, b_i uniformly distributed between $[1 .. K_{max-1}]$, $c_i = 0$)





Analysis:

- *The CAT methods (Bayesian and Adaptive) reduce significatively the number of questions needed to classify a student.*
- *Both methods obtain similar results.*
- *Adaptive seems to be slightly more accurate.*

Simulating an incorrectly calibrated item pool

We assume that for each question there are:

- An ICC_E (k) given by the teacher, defined by parameters a_i, b_i, and c_i in the logistic distribution function.
- An ICC_R (k) that defines the question real behaviour in a test.

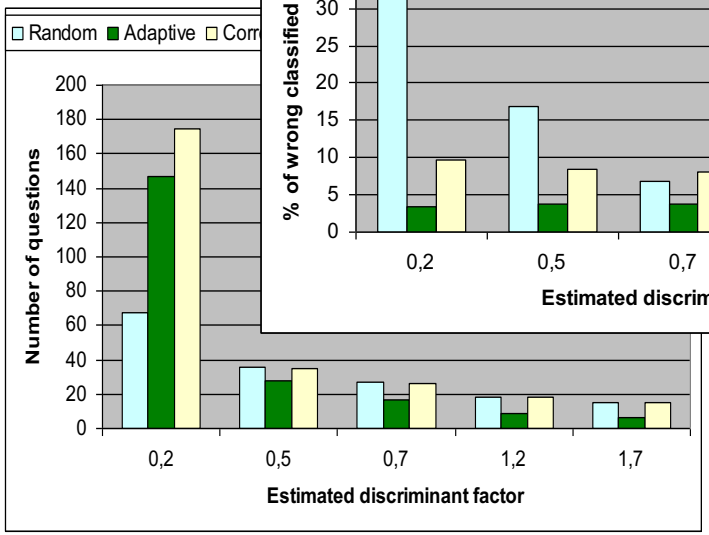
  This function is also assumed to be a logistic function in the simulator

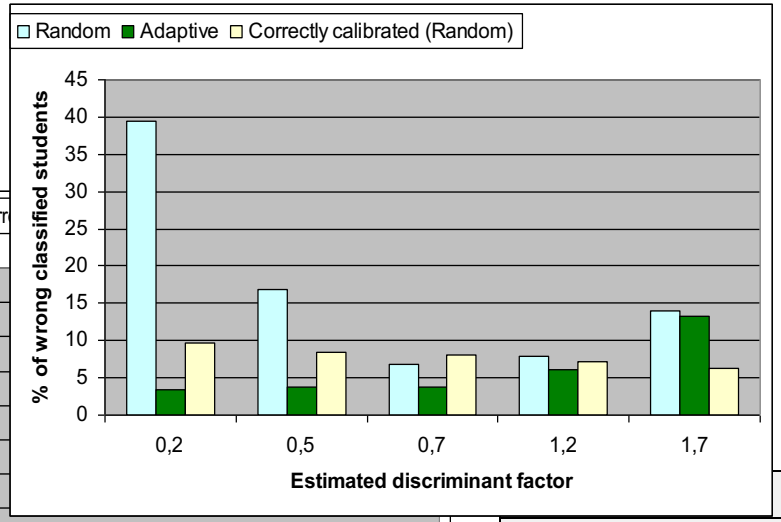
The simulator uses the ICC_R (k) to generate student's responses and ICC_E (k) to simulate the selection criterion, the bayesian evaluation and the finalisation criterion

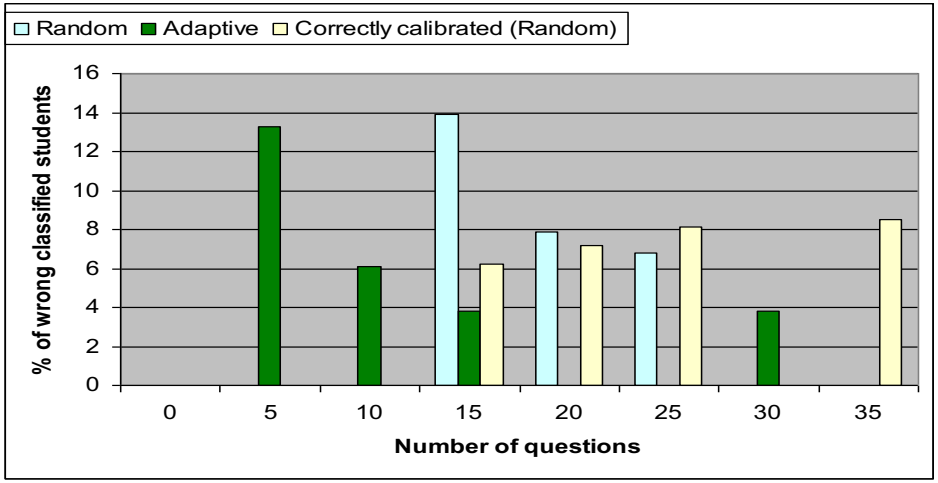
What if the discriminant factor is not correctly estimated? $\rho = 0.90$; $K = 7$

ICC_R (a_i = randomly distributed between 0.7 and 1.7, b_i = uniformly distributed between [1 .. K_{max-1}], $c_i = 0$)

ICC_R (a_i = ... , $b_i = b_i$, $c_i = 0$)







	Random *Selection criterion*		Adaptive *Selection criterion*	
Estimated discrimination factor a_e	*% of correctly classified students*	*Average number of questions posed T*	*% of correctly classified students*	*Average number of questions posed T*
0.2	60.5	67.1	96.6	146.5
0.5	83.2	36.0	96.2	28.0
0.7	93.2	26.6	96.2	16.8
1.2	92.1	18.4	93.9	8.9
1.7	86.1	14.7	86.7	6.4

Analysis:

- *For any "reasonable" estimation of the discrimination factor, the percentage of correctly classified students depends more on the number of questions posed that on the exact value of the estimated discrimination factor*

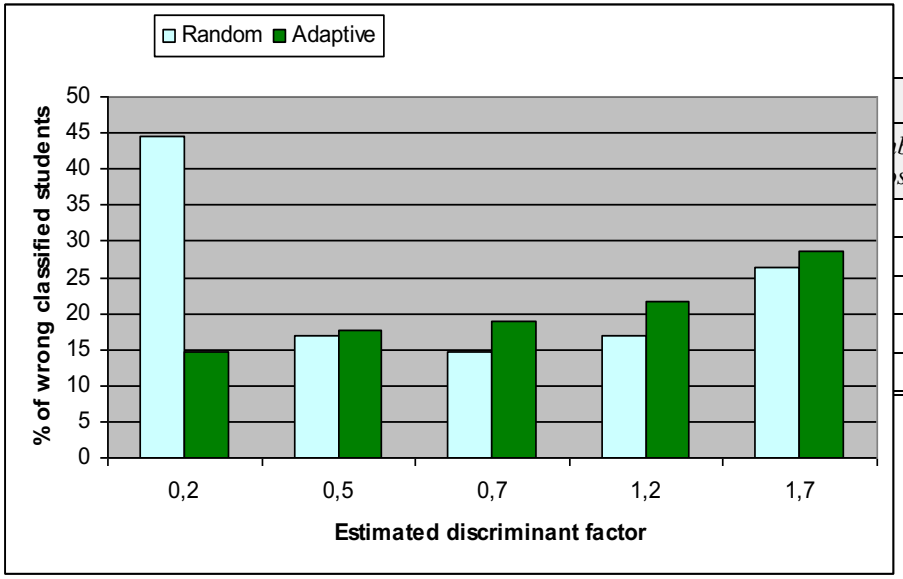
What if the difficulty parameter is not correctly estimated? $\rho = 0.90$; $K = 7$

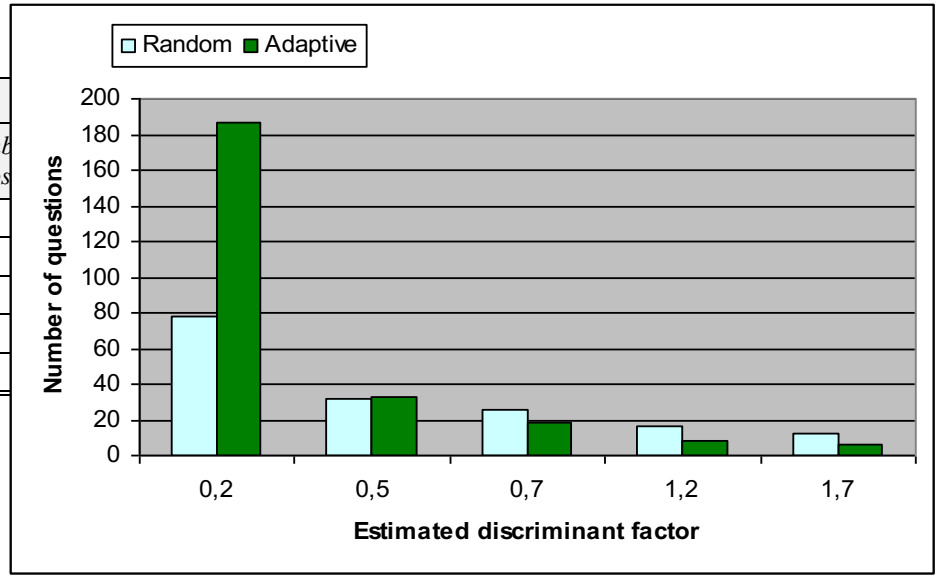
Let us define an *equilibrated item pool* where the questions ICCs are:

ICC_R (a_i = random (~ 1.0) , b_i = uniformly distributed between $[1 .. K_{max-1}]$, c_i)

ICC_E (a_i = ... , $b_i = b_i \pm \lambda_i$, c_i) assuming that the errors are *unbiassed,* that is: $\Sigma \lambda_i = 0$;

Let us consider an equilibrated item pool with 35% of incorrectly estimated questions



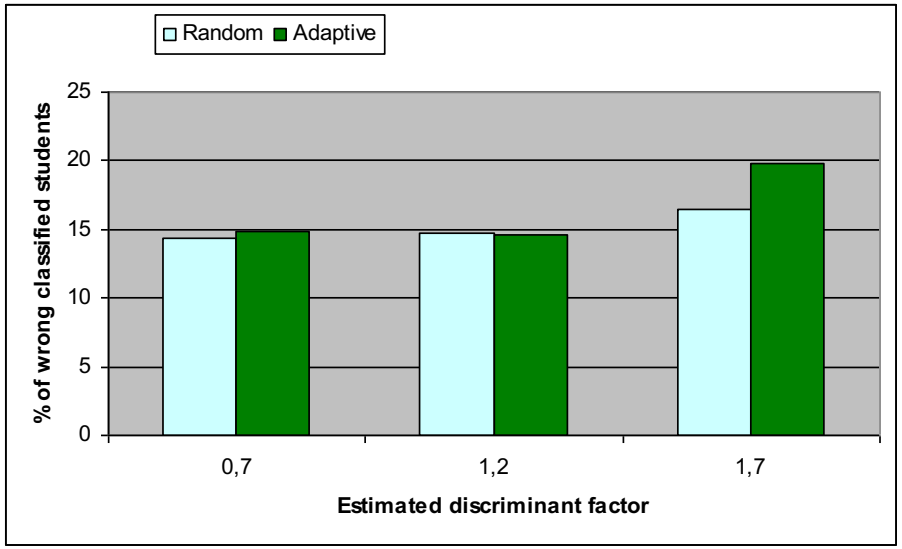


Analysis:

- *Better results are obtained applying the random criterion instead of the adaptive.*
- *The lower the estimated discrimination, the higher the accuracy of the classification.*
- *When the discrimination decreases the number of questions posed increases, and, if it is too small (smaller than 0.5) the accuracy decreases very quickly.*

What if the difficulty parameter is not correctly estimated? $\rho = 0.90$; $K = 7$

Same equilibrated item pool with 35% incorrect questions, considering a fixed number of questions in each test N= 25



Analysis:

- *Similar results are obtained with random and adaptive criterion.*
- *The estimated discriminant factor has not a great influence for central values.*

For each question we define:

•A learned $ICC_L(k)$, that is the ratio between the number of examinees that have answered the question correctly and the total number of examinees that have taken this question and have been classified in knowledge k at the end of a test.

$$\frac{C^+(k)}{C(k)}$$

Measuring the learning

To measure the learning, a distance between questions ICCs is defined:

$$d(ICC_L, ICC_R) = \frac{\sum_{k=0}^{K_{max}} \left| ICC_L(k) - ICC_R(k) \right|}{K}$$

The goodness of the calibration of an item pool can be measured by the average distance among its elements.

Learning modes

Learning takes place when the *current estimated* $ICC_E(k)$ is replaced by the new *learned* $ICC_L(k)$. Learning can be done:

a) *non-incrementally*, that is after a complete set of examinees has passed the test.

b) *by packages,* that is, after a fixed number of examinees has completed the test. (Without keeping the information from previous examinees.)

c) *incrementally*, that is, each time a test is completed and keeping all the information from previous examinees.

In incremental mode a small amount M of experimental cases is included so that initially the learned ICC_L will be equal to the estimated ICC_E

$$ICC_L(k) = \frac{M \times ICC_E(k) + C^+(k)}{M + C(k)}$$

Simulation:

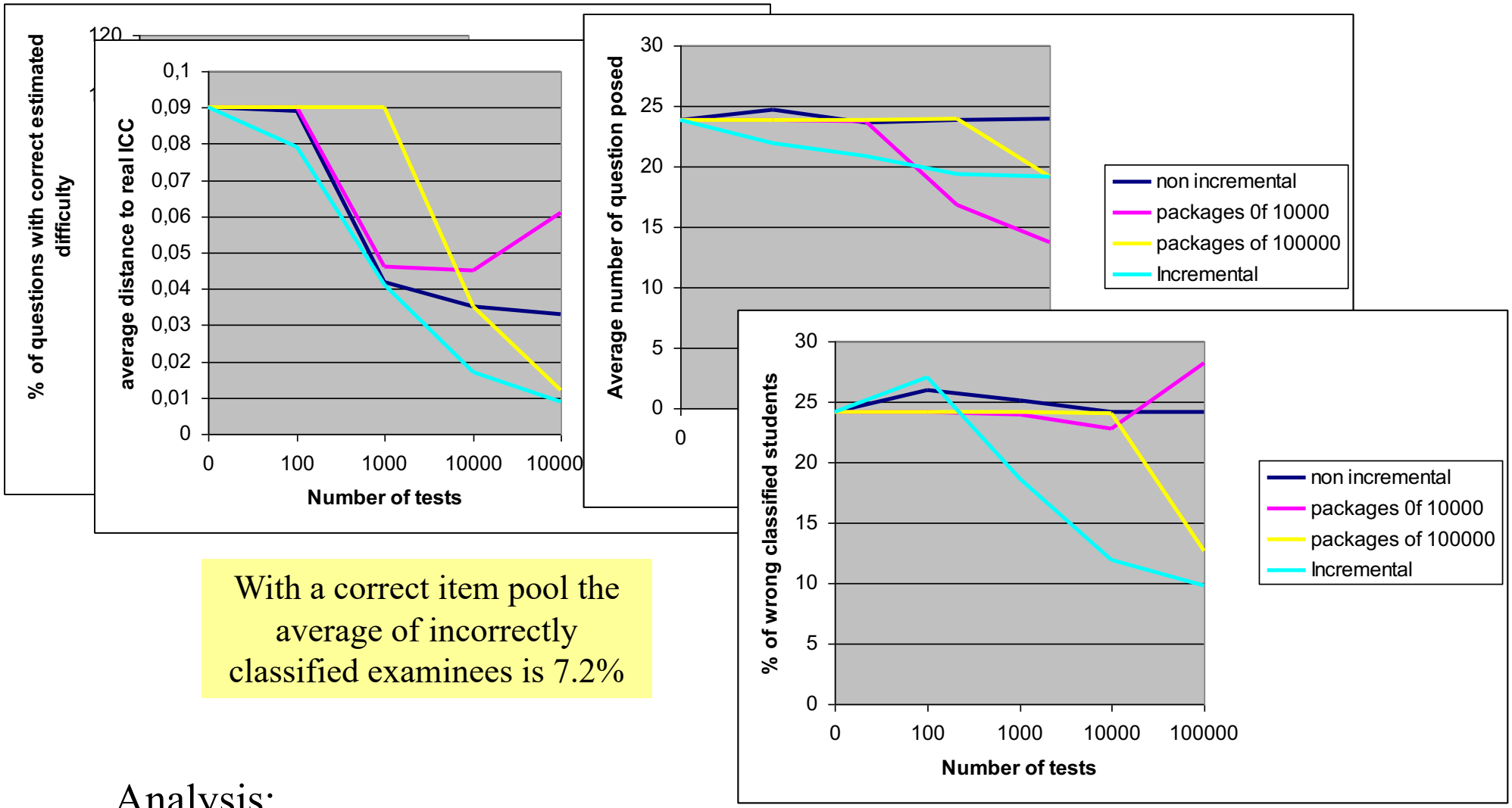
- *Equilibrated item pool* of L = 116 questions, with

  ICC_R (a_i = 1.2 , b_i = uniformly distributed between [1 .. K_{max-1}], c_i)

  ICC_E (a_i = 0.7 , $b_i = b_i \pm \lambda_i$, c_i) where 50% $\lambda_i \neq 0$

- Number of tests: N = 10^2, 10^3, 10^4 and 10^5
- ρ = 0.90, K = 7
- Random selection criterion

Learning procedure	*Examinees learning sample size*	*% of correctly classified students*	*Average number of questions*	*Average cases for learning C(θ)*	*Average distance to the correct set*	*% of questions with correctly estimated difficulty*
Non-incremental learning	0	75.9	23.8	0	0.090	51.7
	100	74.0	24.7	2.8	0.089	49.1
	1000	74.9	23.6	28.9	0.042	94.8
	10000	75.9	23.8	294.6	0.035	100
	100000	75.8	23.9	2945.1	0.033	100
Packages of 1000 learning	0	75.9	23.8	0	0.090	51.7
	1000	76.1	23.7	29.1	0.046	89.7
	10000	77.2	16.8	18.3	0.045	94.8
	100000	71.8	13.7	13.7	0.061	71.5
Packages of 10000 learning	0	75.9	23.8	0	0.090	51.7
	10000	76.0	23.9	293.8	0.035	100
	100000	87.3	19.2	232.3	0.012	100
Incremental learning	0	75.9	23.8	0	0.090	51.7
	100	73.0	21.9	2.1	0.079	58.8
	1000	81.4	20.8	25.2	0.041	94.8
	10000	88.1	19.4	238.4	0.017	100
	100000	90.2	19.1	2360.8	0.009	100



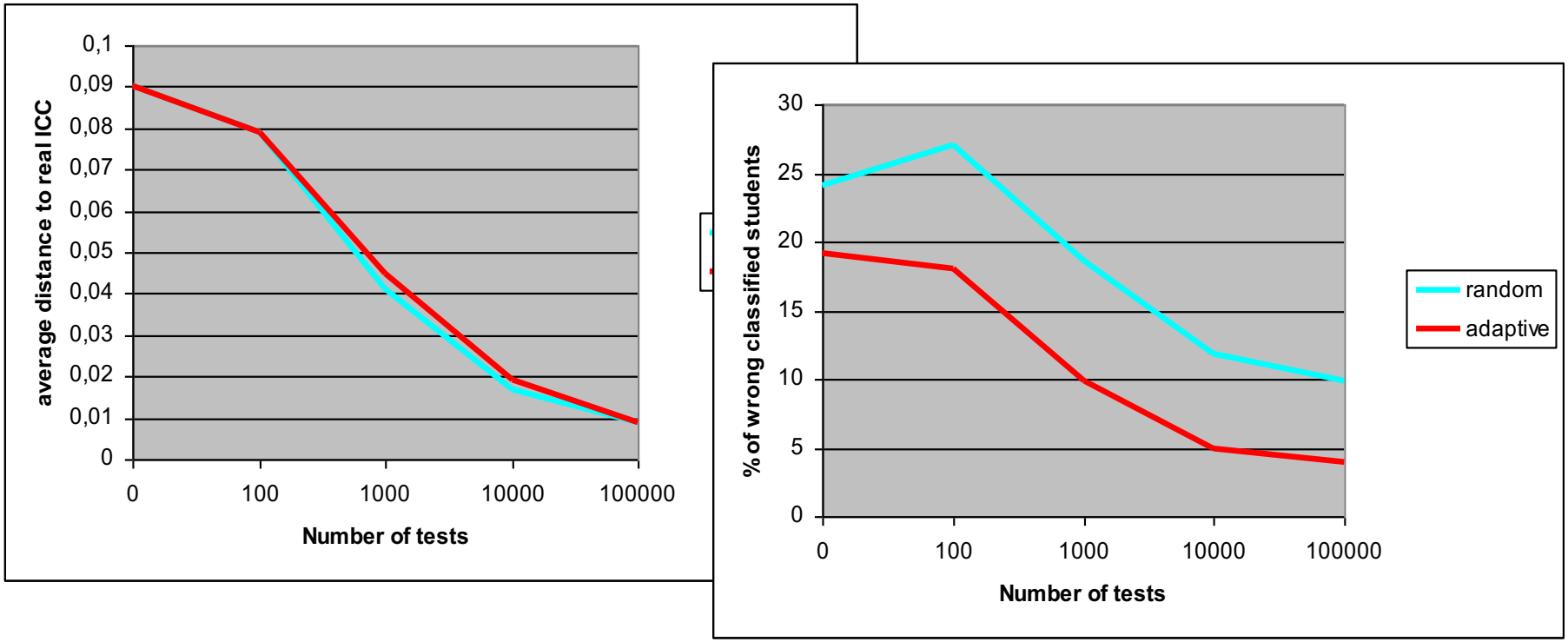
With a correct item pool the average of incorrectly classified examinees is 7.2%

Analysis:

- *Non-incremental learning exhibits good results for approximately more than 10^4 examinees*
- *Package learning is not very good if the package size is smaller than that size.*
- *The incremental learning mode shows the best behaviour.*

Simulation: Same experiment but....

- Fixed number of questions posed in each test: n = 20
- Adaptive selection criterion
- Incremental learning



Analysis:

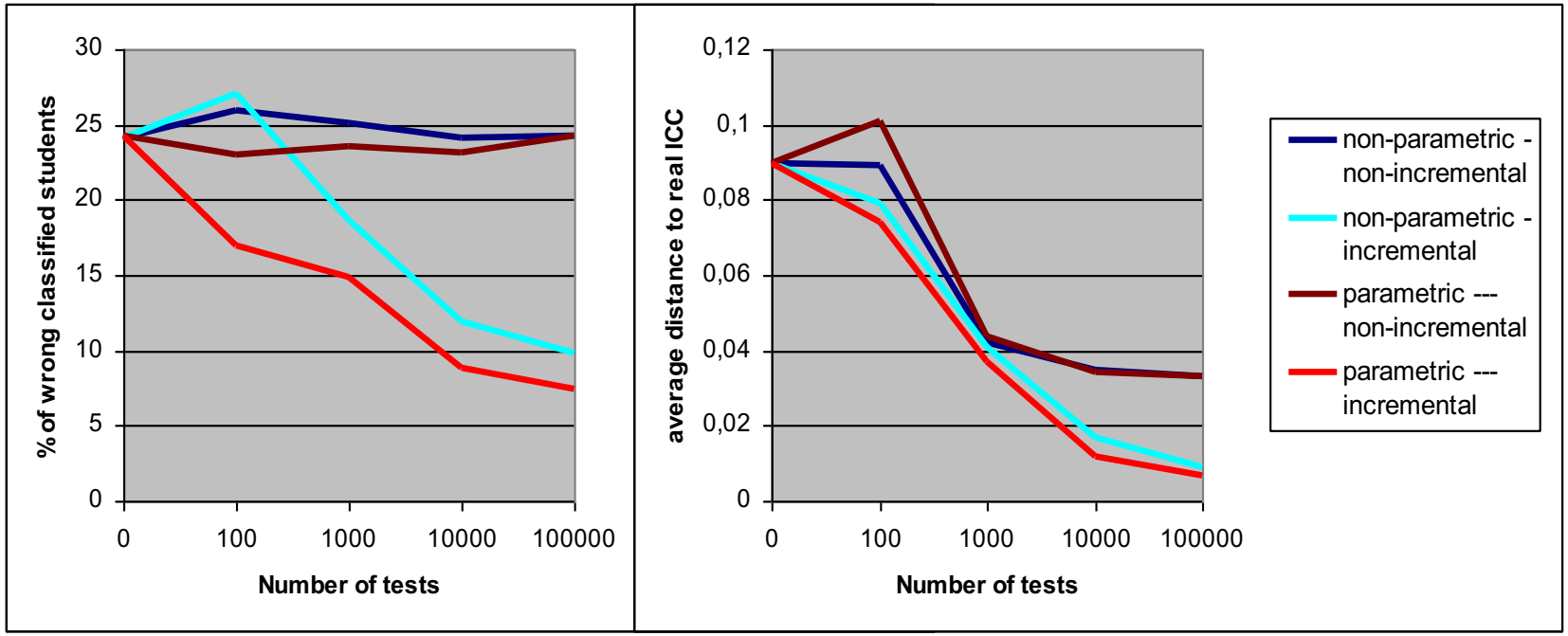
- *The results are now even better than those obtained with random criterion.*
- *With the same number of questions, the adaptive test classifies better than the random test, so learning is also improved.*

Non-parametric vs. parametric learning

• SIETTE is a non-parametric model

• Non-parametric models are more acurate, but they need more information to be calibrated.

•We can convert non-parametic learning to parametric learning by selecting the minimun distance curve from the family of logistic functions.



Analysis:

• *Parametric learning shows better results with a small amount of learning cases.*

• *Results are non conclusive because ICC_R are considered to be logistic*

To clasiffy a student in 5-7 knowledge levels and using adaptive procedure, less than 10 correctly calibrated questions are needed.

- *Adding incorrectly calibrated questions to an existing item pool can reduce its performance when using adaptive criterium.*

- *On-line calibration of the ICCs could be done directly, according to the student's responses and the final result obtained at the end of the test.*

- *If the test is not supposed to be correctly calibrated the best policy to follow is to assign a reasonable low discrimination factor to the incoming questions.*

- *It will also be necessary to turn off the adaptive behaviour or even better, keep the adaptive behaviour but force it to increase the number of questions needed to complete the test.*

- Future work: *theoretical proofs of the results obtained empirically with the simulator.*