An Empirical Approach to On-Line Learning in SIETTE

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

Departamento de Lenguajes y Ciencias de la Computación
Universidad de Málaga. Campus de Teatinos s/n, 29079 Málaga. SPAIN
{conejo, eva, perez, trella}@iaia.lcc.uma.es

Abstract. SIETTE is a web-based evaluation tool that implements CAT theory. With the help of a simulation program, different empirical experiments have been performed with SIETTE with two different goals: a) to study the influence of the parameters of characteristic item curves and selection criteria in test length and accuracy; and b) to study different learning strategies for these parameters. The results of the experiments are shown and interpreted.

1 Introduction

One of the subtasks in an ITS is the evaluation of student's knowledge. SIETTE system [3] has been proposed as a general-purpose web based evaluation system. SIETTE implements *Computer Adaptive Test* (CAT) [5] methodology to improve its performance by reducing the number of questions needed to estimate student's level of knowledge, and is based upon the classical *Item Response Theory* (IRT). SIETTE has been designed as a reusable component to implement a *generic task* [1] for evaluating the knowledge level of a student about certain domain.

Teachers can continuously update the contents of SIETTE question database. This *open architecture* allows the system to evolve and improve its performance over the years. On the other hand, this *on-line* development of question databases is just the opposite of the desired scheme for classical item calibration. Fortunately, the potential great number of students that take the tests provides valuable information that can be used to successively improve teacher's estimations of item parameters.

The main contribution of this paper is an empirical analysis of two issues, namely, the behaviour of SIETTE when using incorrectly calibrated item pools and the feasibility of *on-line* methods for item calibration in SIETTE. The empirical method proposed and implemented uses a program that simulates the behaviour of teachers and students using Monte Carlo techniques.

Item Response Theory (IRT), also known as Latent Trait Theory, was originated in the late 1960s [2]). In a testing context, the *latent trait* is an attribute (*knowledge level*) that accounts for the consistency of test responses. Each question or item is assigned a function (*Item Characteristic Curve*, ICC) that represents the probability of answering to it correctly given the student's knowledge level $\theta \in (-\infty,+\infty)$. Let us represent this probability by the expression: $P(U_i=1|\ \theta)$ or simply by P_i. One of the

G. Gauthier, C. Frasson, K. VanLehn (Eds.): ITS 2000, LNCS 1839, pp. 605-614, 2000.

main problems in IRT theory is to find out the ICCs. It is usually assumed that ICCs belong to a family of functions that depend on one, two or three parameters. These functions are constructed based on the normal or the logistic distribution function. In the three-parameter logistic model the ICC is described by:

$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{1 + e^{-1.7 a_i (\theta - b_i)}}, \qquad \textbf{(1)}$$

where c_i is the guessing factor, b_i is the difficulty of the question and a_i is the discrimination factor. The guessing factor is the probability that a student with no knowledge at all answers the question correctly. The difficulty represents the knowledge level in which the student has equal probability to answer or fail the question, besides the guessing factor. The discrimination factor is proportional to the slope of the curve. If the discrimination factor is high then students with level lower than b will probably fail and students with lever higher than b will probably give the right answer. Assuming that the ICC belongs to this family, the problem of calibrating questions can be formulated as finding the best estimations for the parameters.

Section 2 of this paper describes the implementation of IRT used in SIETTE and the simulator program, and presents some empirical results obtained for correctly and incorrectly calibrated item pools. Section 3 describes a new on-line learning procedure that improves the behaviour of the system by learning item parameters. Finally some conclusions and open issues are addressed.

2 Simulating the Behavior of SIETTE

In this section, we will describe the techniques that we have used to emulate the behavior of the SIETTE system. First, we will describe how to simulate a correctly calibrated item pool.

2.1. Student, Item and Test Simulation

SIETTE implements the IRT model assuming that student's knowledge can be represented as a random variable θ that takes integer values between 0 and K_{max}. This simplification implies that only a fixed and finite number of states of knowledge are considered. Simulated students as proposed in [4] are used. Every student is represented by his/her value for θ. The simulation begins with the random generation of a population of N students, i. e., with the generation of N random concrete values for θ. These values are considered constant during the test (that is, no student learning occurs while taking the test). In the simulations described here the population has been generated to be uniformly distributed in 0, ..., K_{max}. However, other distributions have also been used, not yielding significant differences in the outputs.

Each item is represented by its ICC. An ICCs is also given by K values, corresponding to the conditional probabilities of giving the correct answer to the question given that the student belongs to each of the K classes. The simulator uses a set of Q void questions (ICCs), that are assumed to be correctly calibrated. These ICCs are generated by assigning values to the parameters a, b, and c in a continuous logistic function, and taking the corresponding values for the K percentiles. The simulator allows changing these parameters or to assign them random values, in order to obtain different item pools.

At the beginning of the test, the student is assigned an a-priori probability of belonging to each of the *K* classes in which the students can be classified. The posterior probability is computed applying Bayes' rule. The final result of a test is a distribution of probabilities that the student belongs to each class. The test finishes (in the general case) when the probability of belonging to certain class reaches a fixed threshold ρ (close to 1). This criterion is equivalent to setting a maximum threshold for the standard deviation, which is the one widely used in IRT. Then, we can say that the student belongs to this class with a confidence factor greater than ρ. Other termination criteria can be used, as for example the maximum number of question to be posed.

The simulator successively poses a question to the virtual student and updates his/her probabilities of belonging to each class. This question can be selected randomly or using CATs criteria. The procedure is repeated until the termination criterion is met. Student's behaviour is determined according to his/her estimated knowledge level and the conditional probability that a student of this knowledge level solves the question correctly. That is, if the virtual student has a knowledge level *k* and the value of the question ICC for knowledge *k* is *p*, a semi-random uniformly distributed value *q* in [0,1] is generated. If $q>p$, the system will consider that the student gave a correct answer to the question.

2.2. Simulating a Correctly Calibrated Item Pool

The first empirical analysis carried out concerns how the accuracy of student's classification and the average number of questions posed T depend on the number *K* of knowledge levels considered and on the confidence factor ρ. The percentage of correctly classified students has been computed for an item pool of $Q = 103$ randomly generated questions (ICCs), where *b* is uniformly distributed in $[1, K_{max-1}]$, a=1.2, and c= 0.0 The simulation generates $N = 105$ students. Table 1 shows the results.

The interpretation is that, even with a correctly calibrated item pool, it is not easy to classify "all" the students correctly. This is due to the IRT model itself, that assumes that it is possible (but with a low probability) that a student with a low knowledge level will answer a difficult question correctly and viceversa. The results also show that the percentage of correctly classified students depends more on the confidence factor required that on the number of classes used. On the other hand, the number of questions posed is strongly related to the number of classes considered. For practical reasons, the test should have as few questions as possible, because long tests would be too boring for real students. This practical consideration leads to a compromise between the number of questions and the number of classes.

Table 1. Accuracy of IRT approximation

	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

The second empirical analysis studies how the accuracy of student's classification and the average number of questions posed *T* depend on the quality of the item pool, i.e., on the parameters *a*, *b* and *c*. If *a* increases, the percentage of correctly classified students increases, and the average value of T decreases. If *c* increases, this percentage decreases a little, but the number of questions posed is much bigger. Tables 2 and 3 show the results obtained by using different values for *a* and *c*, (ρ=0.90 and *K*=7).

Table 2. Guessing factor influence

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

Table 3. Discrimination factor influence

Discrimination factor c	*% 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

These results show the great influence of *c* in the number of questions needed. The discrimination factor, *a*, does not have such a great influence in the number of questions if it is bigger than certain threshold. For values smaller than that threshold, the number of questions needed grows very fast. That means that items with low discrimination factor are not informative enough and therefore yield too long tests.

The third empirical analysis carried out concerns how the accuracy of student's classification and the average value of *T* depend on the number *K* of knowledge levels considered and the selection criterion for posing the next question.

It is known that a CAT procedure can be introduced to improve the performance of the classical IRT model. Two different criteria to select the next best question to ask have been implemented in our simulator: a) *bayesian criterion*, that selects the question that minimises the posterior variance of the student knowledge distribution and b) *adaptive criterion*, that selects the question which difficulty equals the average knowledge of the student. Both criteria are equivalent for logistic ICC, as proved theoretically. Table 4 shows the empirical result obtained with the simulator (with ρ=0.90*)*. It is interesting to compare these results with those obtained in the central files of Table 1, that correspond to selecting the items randomly:

Table 4. Accuracy of the CAT approximation

	Bayesian *Selection criterion*		Adaptive *Selection criterion*	
Number of classes K	*% of correctly classified students*	*Average number of questions posed T*	*% of correctly classified students*	*Average number of questions posed T*
3	96.06	3.58	95.62	3.58
5	93.31	6.87	94.67	7.37
7	92.75	8.70	94.43	9.03
9	92.53	9.85	94.23	10.14
11	92.10	10.71	94.14	11.02

The number of questions needed is almost half of the number needed using random selection. These results encourage the use of a CAT procedure, but, as it will be shown later, it is very important to assure that the item pool is correctly calibrated. The adaptive criterion has been chosen over the bayesian one because it gives similar results, but its computational cost is much smaller (this is not surprising, since our ICCs are a discretizations of the logistic model). Similar results are obtained with other discrimination and guessing factors.

2.3. Simulating an Incorrectly Calibrated Item Pool

In Section 3.1, we have assumed that the item pool was correctly calibrated. This is not a fair assumption. In fact it can never be perfectly calibrated, because there is a hazardous component that leads to a known bounded error. To simulate the behaviour of an incorrectly calibrated item pool, let us consider that each question in the database has two ICCs: the *real* ICC and the *estimated* ICC. This is the usual situation when the item pool has been calibrated by a human teacher/expert. Our goal is to study the influence of incorrect calibration in the results of the test. To this end, the simulator uses the real ICC to simulate the answer of the question as described in Section 3.2 and the estimated ICC for any other task.

First, we will assume that the teacher has correctly calibrated the difficulty parameter, but not the discrimination factor a·. Table 5 shows the results obtained assuming that each question has a discrimination factor randomly distributed between 0.7 and 1.7 and that the teacher has assigned a fixed value a_e to all of them (ρ=0.90 and K=7*)*. Compare the results with the ones shown in Tables 1 and 3:

Table 5. Discrimination factor incorrectly estimated

	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

If discrimination factor estimated a_e is bigger than certain lower bound, the percentage of students correctly classified and the number of questions needed do not change very much. 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.

In a second experiment, we assume that some estimations of the difficulty parameter are erroneous, but the error is not biased. That is, sometimes the estimated difficulty is higher and sometimes lower than the real difficulty, but this error is normally distributed around the real difficulty. The same assumption will be made for the discrimination factor. We will call this an *equilibrated item pool*. The justification for this assumption is that, in fact, the knowledge level assigned to a student has not a real meaning by itself: it is only a relative value, like the IQ used in psychology. There is a degree of freedom that is commonly solved in the classical MML parameter estimation procedures by assuming that, for the students in the testing group, the knowledge level has certain distribution. The assumption of an equilibrated item pool reduces this degree of freedom by linking test results to teacher's wishes. If the item pool is prepared by a group of teachers, this hypothesis can be interpreted as a consensus in the meaning of each of the classes (levels) considered. Table 6 shows the results obtained from an equilibrated item pool (randomly constructed) with around 35% wrong assigned difficulty factors, ρ=0.90, and K=7 classes:

Table 6. Equilibrated item pool (ρ=0.90)

	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	55.4	78.2	85.4	186.8
0.5	83.1	32.1	82.4	33.3
0.7	85.4	25.8	81.1	18.3
1.2	83.1	16.0	78.4	8.6
1.7	73.7	12.0	71.4	6.1

Logically, the percentage of correctly classified students has decreased, but the discrimination factor and the selection criterion applied play a very important role. The most significant conclusion is that, if the item pool is incorrectly calibrated, better results are obtained when applying the random criterion instead of the adaptive, which seems very logical. The second is that the lower the estimated discrimination, the higher the accuracy of the classification. Unfortunately, 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.

The good behaviour of small discrimination factors is due to the smaller distance between the estimated and the real ICCs. If the question is incorrectly calibrated, it is better to assume it is not too informative. The fact that the random method shows a better behaviour is explained by the number of questions posed.

In Table 7, the hypotheses are the same as in 6, but we use tests with a fixed number of questions (confidence factor changes accordingly):

Table 7. Equilibrated item pool (fixed number of questions)

	Random *Selection criterion*		Adaptive *Selection criterion*	
Estimated discrimination factor a_e	*% of correctly classified students*	*Average number of questions posed*	*% of correctly classified students*	*Average number of questions posed*
0.7	85.6	25	85.1	25
1.2	85.3	25	85.4	25
1.7	83.6	25	80.2	25

Note that the results are similar (sometimes even better using the random criterion) due to the fact that the main advantage of the adaptive criterion (the smaller number of question it usually needs) was lost when fixing the number of questions. Different results but similar conclusions are obtained with other values for ρ and *K.*

3 On-Line Learning

Taking into account that the results of the test are mainly correct if it can be assumed that the questions set is equilibrated and enough questions are posed to the student; it would be possible to use the results of the test get a better estimation for the ICCs. This has been called *on-line calibration* in IRT literature [5]. None of the methods described for on-line calibration, like the EM or BIMAIN are used in our simulator. However it would be possible to improve the behaviour of the learning mechanism if some extra information could be added, for example if we know that some questions are correctly calibrated and some of them are new (as proposed by Mislevy, cited by Wainer in [5]). A bootstrapping learning procedure can also be used.

In SIETTE, it is possible to learn the probability of each value θ of the ICC array directly from the responses of an examinee that has been classified as belonging to certain class θ. After an examinee has finished a test, all questions that compose the test are fed with the global result obtained and the response (correct/incorrect) to that question. A new *learned ICC* (ICC_L) can be obtained by just dividing the total number of positive cases $C^+(\theta)$ by the total number of cases $C(\theta)$. The better the results of the test, the better the quality of the learning process.

3.1. Incremental and Non-incremental Learning

Learning takes place when the *current estimated ICC* (ICC_E) is replaced by the new *learned ICC* (ICC_L) This could be done a) incrementally, that is each time a test is completed and keeping all the information from previous examinees; b) by packages, that is, after a fixed number of examinees has completed the test. The new ICC is learned only from the most recent examinees' data without previous information; c) non-incrementally, that is after a complete set of examinees has passed the test.

In the incremental and package modes there could be a problem if the number of examinees in the package is small, because some values of the ICC could be out of experimental cases. This problem is even more serious at the beginning of the incremental mechanism, because there is only one case available. The solution to this problem is to include a small amount *M* of initial experimental cases that makes the learned ICC be initially equal to the current estimated ICC. In the simulator, this contour condition has been included only in the incremental mode, so in this case the $\mathrm{ICC_L}$ is obtained by

$$ICC_L(\theta) = \frac{M \times ICC_E(\theta) + C^+(\theta)}{M + C(\theta)} . \qquad \textbf{(1)}$$

In the learning mechanism described below, the number of examinees needed for a calibration depends on the number of questions in the database *L*, the average number of questions in each test $\overline{N}$, the number of classes or knowledge levels that has been considered *K*, and the total number of examinees *n*. The average number of singleton cases that are available to learn the value $\mathrm{ICC_L}(\theta)$ is:

$$\overline{C(\theta)} = \frac{\overline{N}}{L \times K} \times n . \qquad \textbf{(2)}$$

It has to be taken into account that to estimate a probability from *C(θ)* random event observations the following expression applies:

$$p_e = \frac{C^+(\theta) \pm a\sigma}{C(\theta)} , \qquad \textbf{(3)}$$

where p_e is the estimated probability, *p* is the real probability, σ is the standard deviation of the binomial distribution $\sigma = \sqrt{C(\theta) \times p \times (1-p)}$ and *a* is a constant. So, for example to be 95% sure that the real probability is estimated with an error of *p*±0.05, if *p* is in the neighborhood of 0.5 (worst case) we should take a sample of *C(θ)*=400. On the other hand, in this problem not all cases observed come from the right population, because there are also can be errors in the classification process. Our working hypothesis is that the errors present in an equilibrated item pool are compensated. The examinee is sometimes classified higher and sometimes lower.

3.2. Measuring the Learning

The great advantage of using a simulator is that there is complete control over all of the variables that influence the system performance, and that the behaviour of the examinees is only conditioned by their a-priori-known knowledge. So a direct way to measure of the goodness of the learning mechanism could be to measure the improvement in the test performance: the percentage of correctly-classified examinees should increase. Another way of measuring the learning is to define a distance between the real ICC ($\mathrm{ICC_A}$) and the learned ICC ($\mathrm{ICC_L}$). We have selected the simplest distance function:

$$d(\mathrm{ICC_L}, \mathrm{ICC_A}) = \frac{\sum_{k=0}^{K_{max}} \left| \mathrm{ICC_L}(\theta) - \mathrm{ICC_A}(\theta) \right|}{K} \qquad \textbf{(4)}$$

The goodness of the calibration of an item pool can be measured by the average distance among its elements. Table 8 shows the results obtained with each learning mode, at the end of a set of 10^2, 10^3, 10^4 and 10^5 tests, where ρ=0.90, *K*=7, and the initial question database is an equilibrated item pool of *L*=116 questions, with around 50% incorrectly estimated difficulty parameters. The true value for the discrimination factor of all questions in the set is 1.2 but all of them have been estimated initially to be 0.7. The selection criterion was random.

Table 8. Non-incremental, package and incremental learning with random selection

Learning procedure	*Examinees learning sample size*	*% of correctly classified students*	*Average number of questions*	*Average cases for learning $C(\theta)$*	*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

It should be noted that the upper bound of learning is given by the results obtained with a correct set. Table 1a shows that for ρ=0.90 and K=7, the correct set will classify the 92.8% of examinees correctly, requiring an average of 18.1 questions with the random criterion. The percentage of correct classified student shown in Tables 8 are the average during the experiment, including the initial cases when questions have not been modified yet.

Non-incremental learning exhibits good results for approximately more than 10^4 examinees. Package learning is not so good if the package size is smaller than that size. The reason is that there are not enough values to estimate the ICC probabilities for each class. In fact, table 8 show that, if the package is small, there is no convergence. The explanation of this behaviour is that there is a great variance in the learned ICC from just 1000 examinees, and if a poor quality ICC replaces the current estimation the following generation will not be evaluated correctly. Table 8 shows that the incremental learning mode has a better behaviour. ICCs are updated continuously, so both the performance of the test and the quality of the learning process are better. Table 9 shows the results of the same experiment but applying the adaptive criterion to select the question. The criterion to finish the test has been turned off and replaced by a fixed number of questions posed to every examinee, around the same figure that has been used in previous experiment. The results are now even better than those obtained with random criterion. The explanation is that with the same number of questions, the adaptive test classifies better than the random test, so learning is also improved.

Table 9. Incremental learning with adaptive selection and a fixed number of question posed

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 right estimated difficulty*
0	80.9	20	-	0.090	51.7
100	82.0	20	2.0	0.079	55.2
1000	90.2	20	24.2	0.045	94.8
10000	95.1	20	245.9	0.019	100
100000	96.1	20	2462	0.009	100

3.4. Parametric and Non-parametric Models

SIETTE is designed to be a non-parametric IRT model and it is not necessary to assume any shape for the ICCs. Unlike others non-parametric models, SIETTE does not attempt to approximate a continuous function for the ICC from a sparse set of points, but it deals directly with those points. The above learning mechanism does not make any assumption about the shape so it is appropriate for the non-parametric approach. Another point of view could be that SIETTE deals with *K-1* parameters that are the conditional probabilities of each knowledge level. However, there are also some disadvantages in the non-parametric approach. First of all, the classical 1, 2 and 3-parameter models need much less information to be calibrated than the SIETTE model for any *K-1* greater than 3. But a non-parametric learning mechanism can be converted in a more efficient parametric mechanism simply by approximating the just learned ICC by a member of the family of functions considered. This approximation can be done by different methods. In our simulator the sum of weighted minimum squares between the ICC_L and the isomorphic discrete transforms of the logistic family is computed, and the more similar logistic curve is selected. Tables 10 show the results of parametric learning using random selection criterion for 10^3, 10^4 examinees. It should be compared to Table 8.

Table 10 Non-incremental and incremental parametric learning with random selection

	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 right estimated difficulty*
Non incremental parametric learning	0	75.8	23.9	-	0.090	51.7
	100	77.0	22.2	2.1	0.101	44.8
	1000	76.5	23.4	27.4	0.044	92.2
	10000	76.9	23.9	293.4	0.034	100
	100000	75.8	23.9	2949.5	0.033	100
Incremental parametric learning	0	75.8	23.9	-	0.090	51.7
	100	83.0	25.8	2.8	0.074	60.3
	1000	85.1	22.4	27.3	0.037	96.5
	10000	91.2	20.7	255.6	0.012	100
	100000	92.5	20.3	2509.8	0.007	100

Another interesting point is that there seems to be a limit in the approximation that can be achieved with non-incremental learning, either parametric or non-parametric. The explanation of this residual error probably lies on the variance of the random selection of questions from the equilibrated set. The original 116 question set is equilibrated, but the subsets of questions used in the test are not necessary equilibrated.

4 Conclusions

Using a simulator program and applying Monte Carlo methods, we have studied the behaviour of IRT and CAT in the SIETTE system in order to know the quality of the information that can be extracted from a single test and the expected number of questions needed.

For most applications in ITS it is enough to deal with 5-7 knowledge levels about the domain. Less than 10 questions are needed (if they are correctly calibrated). It is desirable to initially calibrate the question set, but it is also possible to trust in the criterion of the teacher(s) that defines the test, and improve its performance by the on-line learning mechanism described. On-line calibration of the ICCs could be done directly, according to the responses of the student and the final result obtained at the end of the test. It also can be done more efficiently if it can be assumed that the ICCs shapes can be described by a family of functions.

If the test is not supposed to be correctly calibrated (i.e. many new questions have been added recently) 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. This constraint should be eliminated once the question set has been self-calibrated.

The results presented in this paper are obtained from empirical experiments in a simulated environment. It would be also necessary to develop some experiments with real world data. On the other hand, we are currently working in a formalisation of the concept of *equilibrated item pool* and in a theoretical demonstration of the results obtained empirically with the simulator.

References

1. Chandrasekaran, B. (1992). Generic Task: Evolution of an idea. *Technical Report. Laboratory for AI Research. Ohio University. Columbus OH.*
2. Lord, F. M. &. N. M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
3. Ríos, A., Millán, E., Trella, M., Pérez-de-la-Cruz, J. L., & Conejo, R. (1999). Internet Based Evaluation System. In *Proceedings of the 9th World Conference of Artificial Intelligence and Education AIED'99* (pp. 387-394).
4. VanLehn, K., Ohlsson, S., & Nason, R. (1995). Applications of Simulated Students: An Exploration. *Journal of Artificial Intelligence and Education, 5*(2), 135-175.
5. Wainer, H. (ed.). (1990). *Computerized adaptive testing: a primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.