AN ADAPTIVE ASSESSMENT SYSTEM TO EVALUATE STUDENT ABILITY LEVEL

Antonella Carbonaro, Giorgio Casadei, Simone Riccucci
University of Bologna, Italy

Abstract: The experience from years of development and use, the advance of technology, and the development of authoring tools for questions and tests has resulted in a sophisticated, computer based assessment system. However, there is still a lot of room for further development. Some of the current ideas for development are discussed in the remainder of this work. A primary aim of assessment, both formative and summative, is provide the necessary information to improve future educational experiences because it provides feedback on whether the course and learning objectives have been achieved to satisfactory level. Yet, it is important that the assessment data be accurate and relevant to effectively make informed decisions about the curriculum. Moreover, formative assessment can also be used to help bridge the gap between assessment and learning. This may be achieved particularly where assessment strategies are combined with useful feedback, and integrated within the learning process. The answers to the described objectives are enhanced if we could integrate adaptive testing techniques; accurate and fitted assessment data may improve both the curriculum and the student ability level. The idea behind a computerized adaptive testing (CAT) is quite forward: to apply to each examinee only those items useful to know his proficiency level. As a consequence of this, CAT is more efficient than conventional (i.e., fixed-item) tests. It provides more precise measurements for same-length tests or shorter tests for same-precision measurements.

Key words: Assessment, Item Response Theory, Computerized Adaptive Testing, Ability

1. SYSTEM DESCRIPTION

The systems used to deliver our tests, have approximately the same functions even if they have been developed in two different architectures. The first one, “Examiner”, was developed using ASP (Active Server Pages)

technologies and Access as Database Management System, the second one, "XTest", following the experience of the first system, was developed in Java/JSP (Java Server Pages) and mySQL as DBMS to make it platform independent. The task performed by these systems is to create and manage questions of various types, create randomized test for an exam given some constraints on contents, deliver the test created to a client application. For Examiner, the questions are delivered as HTML text to a web browser that render them on the screen, while for XTest a Java Applet receives data from a server application launched from a machine that acts as control console. The main advantage of first solution is the minimal requirements of client system resources so that only a browser HTML 1.0 compatible is needed to run the client application. The second solution need to install a JRE (Java Runtime Environment) on the client machine, but a Java Applet has more potential in creating new questions type and in a large distributed environments, it allows to have less computational loading on the server.

In such systems, the teacher has to create and insert some question that is grouped by their subject topics and in a successive phase, they will decide how to build the test for the session. The courses concern base competences on Computer Science and Prolog programming. The courses have been divided in six topics:

- GLOSSARY: the questions belonging to this topic, ask for the meaning of some word or the functions of some objects of the computer world;
- FOUNDAMENTALS: this topic concerns basic knowledge on calculability, algorithms, complexity, computer architecture, compiler and programming languages;
- PROLOG: question that ask to interpret or complete some pieces of Prolog source code;
- PROLOG01: this topic is about the Prolog syntax;
- PROLOG02: this topic is about the problem formalization in prolog;
- TURING: exercises on turing machine;

The item type chosen for student assessment was closed answer type. In particular, even if the systems are able to manage more than one item type, the multiple response questions have been used. This kind of question consists in a text representing the question itself and a list of n possible answers that the examinee has to check if the answer is right. For each answer is associated a score that is positive if the answer is right and negative otherwise. A zero score is assigned for not checked answer. The sum of the single score, gives the exercise result. The questions are sequentially presented to the student and if they are checked, the student is not allowed to review question. Otherwise, he can review it once again.

Furthermore, the questions are randomly presented to give the feeling that each test is different from each other.

2. THEORETICAL FUNDAMENTS

Item Response Theory is used as mathematical model providing necessary framework to our system. In this measurement theory, there is an attempt to relate some unobservable characteristics, like getting good grades, learning new material easily, relating various sources of information, and using study time effectively, to observable variables like test results. For the purpose of this work, we assume that there an underlying arbitrary characteristic associated to the examinee that we call ability. This is the unidimensional assumption, that is only one kind of unobservable characteristic is needed to complete the test.

The relation between item examinee performance and his ability can be expressed by a monotonic increasing function called Item Response Function or Item Characteristic Curve.

Such a function states that the probability of correct answer grows as the underlying ability grows. Starting from these assumptions we can mathematically model our system, with this theory.

Each model has its own set of parameters (constants) to associate with item. Typical parameters are difficulty level and a discrimination power. The main advantage of this theory, is that the scale to measure ability is unique regardless both the item and person sample.

It is possible to estimate such parameters starting from a sample of responses given in real exam sessions, through a procedure called EM algorithm, so we can associate the parameters to the items and setup the adaptive system.

2.1 Item Information Function

Item Information Function (IIF) is a powerful tool to evaluate and construct a test, either in fixed or adaptive way. Intuitively this function tells us the amount of information given by an item about an examinee with a given ability. This function evaluates the utility of an item to estimate an examinee that is supposed to have certain ability. Analyzing the function shape, in particular its peak, we can have a visual appealing of the item difficulty.

As an example, the more the difficulty level is closer to the examinee ability, the bigger IIF value is (it is desirable to assess high skill students with very difficult question and low prepared students with easy question).

IIF allow us to estimate the error made in evaluating examinee ability by computing its inverse square root.

2.2 Generalized Partial Credit Model

The choice of model for our system is the Generalized Partial Credit Model [MUR1992] because of the multi-category nature of the items belonging to item bank. In this IRT model a score $X_i = 0, 1, .., m_i$ can be obtained on item i =1,., B from an item bank with B items. A higher score indicates a better performance and m_i indicates the maximum score on item i.

The probability of obtaining a score k on item i, given the value of the ability θ, is denoted by

$$P_{ik}(\theta) = \frac{e^{\sum_{v=0}^{k} a_i(\theta - b_{iv})}}{\sum_{c=0}^{m_i} e^{\sum_{v=0}^{c} a_i(\theta - b_{iv})}} \qquad k = 0,1,\ldots,m_i \;\; b_{i0}=0$$

where a_i is the discrimination parameter and b_{iv} is the item step parameters.

Note that instead of the ICC, here we have a set of m_i *Item Category Response Function* that correlates the ability level with the category response probability.

The IIF for this model is given by the following equation

$$I_i(\theta) = a_i^2 \left[\sum_{k=0}^{m_i} k^2 P_{ik}(\theta) - \left(\sum_{k=0}^{m_i} k \; P_{ik}(\theta) \right)^2 \right]$$

2.3 Computerized adaptive testing

The base idea of this assessment technique, is that during the test, the ability of an examinee can be estimated and the useful item for the current ability be chosen from item bank (generally the item that will mostly reduce the ability estimation error at the test end).

Testing process can be divided in the following phases:

1 – How to start a test:

if we have no information about the examinee, we suppose that he has a medium ability usually corresponding to value 0. Upon this consideration the system chooses the most suitable item for the ability 0. Another strategy

is to deliver two or three items randomly chosen in order to obtain initial ability estimation.

2 – How to continue a test:

many strategies can be applied to item selection. The most common is based on Item Information Function (IIF). As mentioned before, this function represents the amount of item information at a certain point on the ability scale. The more high information is, the more accurate will be the ability estimation. The item selection algorithm based on this function, at each step of the test, estimates the examinee ability and gets the most informative item, for such ability, from the item bank. In this way it attempts to reduce at minimum the estimation error.

3 – How to stop a test:

the stopping rule is based on three conditions: reaching of prefixed estimation precision, maximum test length and maximum time to finish the test.

The selection algorithm described above, it's not efficient in terms of item exposure rate. In fact, for some reasons due to the IRT, it always tends to choose the best quality items in terms of discrimination power (as mentioned above more discrimination power correspond to more information given by an item), making necessary the need of some control mechanisms. The adopted mechanism for our system are the "SH algorithm" for overexposure rate control and the "progressive method for exposure control" [EGG2001] for underexposure control. The system also need for a content balancing mechanism, due to the need of specify the test content. The "Constrained CAT" algorithm [LCH2001] was implemented.

2.4 System setup

To make operative our system, we need for item parameters estimate, from real data collected during exam sessions. Data used was collected in various exam sessions (about 1400 tests done) of Italian Bologna University in first year undergraduate courses of faculty "Economia di internet" in Forlì department, "Economia del turismo" in Rimini and finally "Psicologia" in Cesena.

To make operative this system, we have collected and analyzed the exam results we have done in the following way:

- the items are in Multi Response form, where to each answer is associated a score stating the correctness level. The final score is expressed by the selected answer score sum.
- each question is allowed to be reviewed once and it is not possible to change the answer once the student confirms it. A time limit was imposed depending on the number of item.

Now we have a minimum and a maximum score for each question (item) answered by the examinees. We need to transform this data in a format according to Generalized Partial Credit Model. The choice was to model the items with a five category GPCM by dividing the items score range in five intervals. Thus for the first interval we have assigned the value 0, for the second the value 1 and so on. The data obtained is a matrix of integer number where in each row we have the result of an exam and in each column the result for each question. Once we have done this transformation, we have used ICLWin software [HAN2002] for the item calibration phase.

The adaptive approach for testing can now be used taking advantage of IRT paradigm (Computerized Adaptive Testing [WAI2002]). During the test the ability of examinee can be estimated and the useful item for the current ability be chosen from item bank (generally the item that will mostly reduce the ability estimation error at the test end).

3. SIMULATION STUDIES AND RESULTS

To evaluate the system efficiency, in term of estimation error and item exposure rate, we have simulated some exam sessions with the adaptive method and compared the results with the fixed test mode used upon now.

In real exam sessions, teacher chooses which arguments to include in the test and how much items for each argument include in the test and the system gets the items randomly from the item bank, with respect to specified topic constraint.

For simulation studies, we assume that the number of items is equally distributed around the arguments and the arguments are randomly chosen for every simulated session. For each simulation run, the system generates 200 exam session composed by n examinees with n from a uniform distribution in the range [30-200], so we have about 23000 examinees for each simulation. Each examinee has associated an ability value from a Gaussian normal distribution with mean 0 and variance 1. With a random mechanism, k arguments are chosen for each session. An additional constraint is that a minimum length for a test is 10 items. The stop rules for adaptive test is a maximum length of 30 items and a maximum error of 0.3 on ability scale.

First simulation was conducted by turning off all the control on content balancing and exposure to evaluate the need of such control mechanism. From the first simulation, we have obtained that item exposure rate is quite uniform across the entire item pool and the maximum exposure rate is lower than 8%. This is an acceptable upper bound for exposure rate so the overexposure control will be disabled in next simulation. There will be the need for underexposure control, because some of the item has not been used

at all. Figures 1-3 show the result of a more realistic situation, where both content balancing and underexposure rate control is applied. The figures show a substantial enhancement in item exposure rate with an acceptable loss of performance. The test length is increased of about five items, on average, while the estimation error is almost the same. To complete the system evaluation, we now compare the estimation error made with the past real exams, with that of computerized adaptive testing system developed.

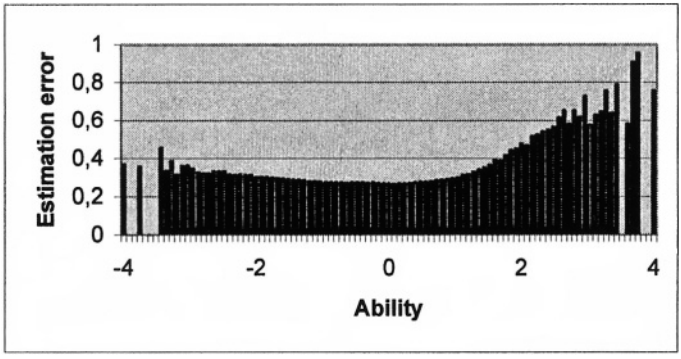

Figure 1. Estimation error as the ability function

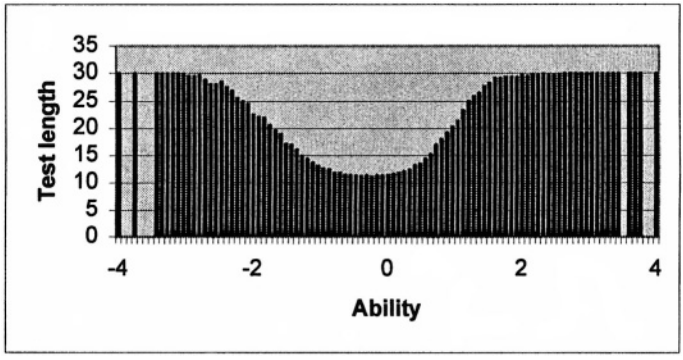

Figure 2. Test length as ability function

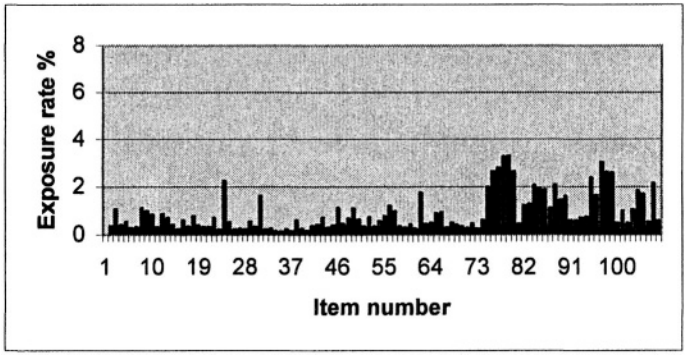

Figure 3. Item exposure rate for each item

Figure 4 shows a scattered graphic where for each exam *i* we draw a point (θ_i, Err) where θ is the estimated ability and Err is the estimation error for the ability. The administered tests are 30 items long.

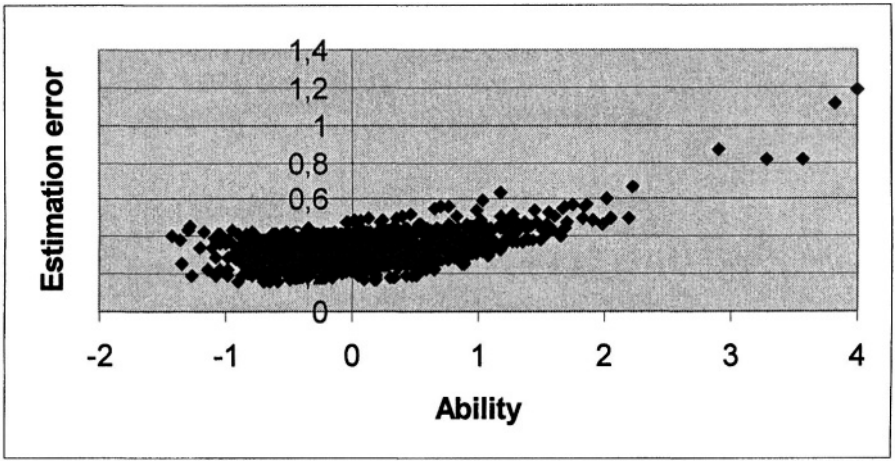

Figure 4. Scatter chart of estimation error for real exams

4. CONCLUSION AND FUTURE WORK

We have seen that the IRT framework gives us a powerful tool to evaluate student proficiency and it allow to build, on solid mathematical bases, a dynamic system that understand the ability level during the test delivery phase.

Computerized adaptive testing is about twice more efficient than the fixed-item test and a further development is needed. On the other side, this kind of system needs a relatively complex process of setup (pre-test data collection, item evaluation and calibration). Next step of our work is to make the system capable to learn from its experiences, automating this phase.

Very interesting is to introduce the new IRT models that include more than one characteristic to evaluate. Such a models are object of intensive research and in the next few years, they can be utilized to develop systems that are more complex.

Another important problem is the need for very large amount of data in calibration phase to make item parameters reliable. The introduction of artificial intelligence techniques like neural networks should be useful both in the calibration process and to help manipulating more complicated models.

REFERENCES

[EGG2001] T.J.H.M. Eggen, "Overexposure and underexposure of items in computerized adaptive testing", 2001, http://download.citogroep.nl/pub/pok/reports/Report01-01.pdf

[HAN2002] Bradley A. Hanson, IRT Command Language (ICL), 2002, http://www.b-a-h.com/software/irt/icl/

[LCH2001] Chi-keung Leung, Hua-hua Chang, Kit-tai Hau, "Making *a*-Stratified Computerized Adaptive Testing Design More Practical: Imposing Non-statistical Constraints",2001,http://www.fed.cuhk.edu.hk/GCJCE/gcjce04/gcjce04.html

[HSR1991] Hambleton R.,Swaminathan H., Rogers J., *Fundamentals of Item Response Theory,* NewBury Park, SAGE Publications, 1991.

[MUR1992] Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159-176.

[REHS2000] P.W. van Rijn ,T.J.H.M. Eggen ,B.T. Hemker ,P.F. Sanders ,"A Selection Procedure for Polytomous Items in Computerized Adaptive Testing", 2000, http://download.citogroep.nl/pub/pok/reports/Report00-05.pdf

[WAI2000] Wainer H. and other, *Computerized Adaptive Testing: A Primer,* 2° edizione, Mahwah, Lawerence Erlbaum Associates, 2000.