# COMPUTER-AUTOMATED TESTING: AN EVALUATION OF STUDENT PERFORMANCE

Silvia Cagnone, Stefania Mignani and Roberto Ricci
*University of Bologna*
*Statistics Department*
{cagnone, mignani, rricci}@stat.unibo.it

Giorgio Casadei and Simone Riccucci
*University of Bologna*
*Computer Science Department*
{casadei, riccucci}@cs.unibo.it

**Abstract** In the last few years, the need for an automated way to assess people has increased quickly because of the growing request from both private and public structures. Many Learning Management Systems (LMS) have been developed in order to automatize the learning and assessment process. In most of the cases these systems don't allow a quality content evaluation and an efficient ability estimation. In this paper we analyze the features of the Proportional Odd Model (POM), belonging to the Item Response Theory. The POM enables to translate an automatic test deliver in an efficient way in order to get an evaluation through either a summative or a formative way. The data have been collected in some undergraduate courses of Bologna University, by using test delivering and by authoring system developed in ASP and Java, respectively.

**Keywords:** Automatic test deliver, automatic evaluation system, undergraduate students ability system, IRT models for ordered polytomous variables.

## Introduction

In the Italian educational system the increasing level of formative requirement needs a particular consideration in the assessment and evaluation field. The assessment, defined as process of measuring learning, is a problematic component of the most e-learning programs. Each automatic evaluation system requires the introduction of methodological statistical tools.
Within a new way of understanding the assessment process of a student, eval-

uation may serve two complementary functions. In one context, the aim is prospective, or *formative* - to improve, to understand strengths in order to amplify them, or to isolate weaknesses to mend. Formative evaluation is a process of ongoing feedback on performance.
The other context is retrospective, or *summative* - to assess concrete achievement, perhaps as part of a process of acknowledgement or giving awards. Summative evaluation is a process of identifying larger patterns and trends in performance and judging these summary statements against criteria to obtain performance ratings.
Before new methodologies realize their fullest potential we must expand our basic mental model of what they are [Gentner and Stevens, 1983]. Cognitive psychologists define mental models as the way to understand and analyze a phenomenon. That is, people have different mental models of e-learning, depending on their attitude to it and their experiences with it. It's very important to focus our attention on the assessment problem of a examinee performance. We have to conceive it like the exterior expression of a set of latent abilities. The statistical evaluation of these abilities and its transformation into a mark are the principal aims of this work.
In our particular case we consider an experimental project developed at the University of Bologna based on an automatic evaluation system applied in different steps of the educational offer.
In this paper we intend to investigate the efficacy of a test delivering system that automatizes the task to submit and correct a test. In a computer-based testing, there are many issues to be considered: test administration, the impact that the system will have on examinees and the way to assign a final mark.

## 1. The 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", has been developed by using ASP (Active Server Pages) technologies and Access as Database Management System. The second one, "XTest" has been developed, following the experience of the first system, in Java/JSP (Java Server Pages) and mySQL as DBMS to make it platform independent.
The task performed by these systems is to manage questions of various type, by creating randomized tests for an exam given some constraints on contents and by delivering them to a client application. For Examiner, the questions are delivered as HTML text to a web browser that renders them on the screen, whereas for XTest there is a Java Applet that receives data from a server application launched from a machine. The machine acts as control console. The main advantage of the first solution is the minimal requirements of client system resources so that only a compatible browser HTML 1.0 is needed in order

to run the client application. The second solution needs a JRE (Java Runtime Environment) to be installed on the client machine. A Java Applet has more potential in creating new questions type and in a large distributed environments. Furthermore, it allows to have less computational loading on the server.
In such a systems the teacher has to create and insert some questions that are grouped by their subject topics. In a successive phase, he was to decide how to build the test for the session.
Each item is shown to the student through a window in which he has to put his answer. A typical case is an item that displays a Prolog algorithm selecting all negative numbers from a list of integers and inserting them to on output list. The examinee has to understand what the Prolog code does and to select the interpreter output. The examinee has to understand what the *Prolog* code does and has to select the interpreter output.

The courses concern basic competencies on Computer Science and Prolog programming. The courses have been divided in five topics:

**GLOSSARY** : the questions belonging to this topic ask for the meaning of some words or the functions of some objects of the computer world;

**FOUNDATIONS** : this topic concerns basic knowledge on calculability, algorithms complexity, computer architecture, compiler and programming languages;

**PROLOG** : question that asks to interpret or complete some parts of Prolog source code;

**PROLOG01** : this topic is about the Prolog syntax;

**PROLOG02** : this topic is about the problems formalization in Prolog.

The item type chosen for student assessment is a 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 of a text representing the question itself and of a list of $n$ possible answers that the examinee has to check if the answer is right. A score is associated to each answer. It is positive if the answer is right and negative otherwise. A zero score is assigned to not checked answer. The sum of the single scores gives the exercise result.
The questions are presented to the student sequentially and if they are checked, the student is not allowed to review them. Otherwise he can review it once again. Furthermore the questions are randomly presented so that each test seems to be different from each other.

## 2. The statistical evaluation: the ability as latent trait

Usually the analysis of the results of a test is not taken into account independently from the formulation of the questionnaire. In the classical test theory, each item is evaluated through a score and the *total score* permits to give a mark to the examinee.
At the beginning of the sixieties a new methodology, called *Item Response Theory* (IRT) [Lord and Novick, 1968] has been developed: it allows to evaluate the student ability, the question difficulty and the capability of the item to distinguish between examinees with different ability. Since ability is not directly observable and measurable, it is referred to as latent trait. Thus an IRT model specifies a relationship between the observable examinee text performance and the unobservable latent trait (ability) that is assumed to underlie the test result. A mathematical function permits to describe the relationship between the "observable" and the "unobservable" quantities. This function is called *Item Characteristic Curve* (ICC) and it determines the probability to answer correctly to an item on the basis of a given ability examinee level, through the ICC.
The ICC is described by parameters (constants) whose number depends on the specific chosen IRT model. In general one constant represents the *difficulty* of the item, that is, a high value of the parameter implies a high level of complexity. An other parameter, called *discrimination power*, is used to quantify the capability of the question to distinguish between examinees with different ability level.
After the estimation of the parameters it is possible to determine the examinee ability that is transformed into a mark by a function illustrated below (see par. 4.3).
Unlike classical test models the IRT presents different advantages, but the most important is that it allows to estimate a student ability on the same ability scale from any subset in the domain of items that have been fitted to the model [Hambleton and Swaminathan, 1985]. An ability estimation which is independent of the number and the choice of items represents one of the most important advantages of IRT models. They provide a way of comparing students even though they have taken different subset of test items. Assuming a quite large population of examinees, the parameters of an item are independent of the particular sample of students.

## 3. Model specification

In this work we apply a particular IRT model called Proportional Odds Model (POM) introduced with the aimto treat the case of ordinal observed variables. Problems with different levels of complexity have been included in each argument (item). In fact, the problem solving process contains a finite number of steps so that the student ability can be evaluated on the basis of the step

achieved. In this way, for the $i$-th item an ordinal score $m_i$ is assigned to the examinee who successfully completes up to step $m_i$ but fails to complete the step $m_i + 1$. Following this procedure, a score ranging from 1 to 4 is assigned to each examinee for each item with respect to the solving level achieved (1=no correct answers, 2=correct answers only for preliminary problems, 3=correct answers also for intermediate problems, 4=all correct answers).
Let $x_1, x_2, \ldots, x_p$ be $p$ ordinal observed variables regarding the items of a questionnaire and let $m_i$ denote the number of categories for the $i$-th variable. The $m_i$ ordered categories have probabilities $\pi_{i1}(z), \pi_{i2}(z), \ldots, \pi_{im_i}(z)$, which are function of $z$, the latent trait representing the individual ability. They are known as *category response functions*. Indicating with $\mathbf{x}_r = (x_1, x_2, \ldots, x_p)$ the complete response pattern of the $r$-th individual examined, we can define the probability of $\mathbf{x}_r$ as:

$$\pi_r = \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} \pi_r(z) h(z) dz \tag{1}$$

where $h(z)$ is assumed to be a standard normal and $\pi_r(z)$ is the conditional probability $g(\mathbf{x}|z)$. For $g$ the conditional independence is assumed, that is when the latent variable is held fixed, the $p$ observed are independent. In the case of ordinal observed variables it is defined as:

$$g(\mathbf{x}_i|z) = \pi_r(z) = \prod_{s=1}^{m_i} \pi_{is}(z)^{x_{is}} = \prod_{s=1}^{m_i} (\gamma_{is} - \gamma_{is-1})^{x_{is}} \tag{2}$$

where $x_{is} = 1$ if a randomly selected person responds in category $s$ of the $i$-th item and $x_{is} = 0$ otherwise and $\gamma_{is} = \pi_{i1}(z) + \pi_{i2}(z) + \ldots + \pi_{is}(z)$ is the probability of a response in category $s$ or lower on the variable $i$. $\gamma_{is}$ is know as *cumulative response function*. The model is defined in terms of a logit function of $\gamma_{is}$ and can be expressed in a general form within the generalized linear models framework as Moustaki (2000):

$$logit(\gamma) = \ln \left[ \frac{\gamma_{is}(z)}{1 - \gamma_{is}(z)} \right] = \alpha_{is} - \beta_i z \quad , \qquad s = 1, 2, \ldots, m_i - 1 \tag{3}$$

The model so defined is called Proportional Odds Model (POM) and is very similar to the well known Graded Response Model by Samejima (1969). It ensures that the higher is the value of an individual on the latent variable, the higher is the probability that individual belongs to a higher item categories. The intercept parameter $\alpha_{is}$ can be interpreted as the item *difficulty* parameter whereas $\beta_{ij}$ can be interpreted as the *discrimination* power parameter. To define ordinality properly, the condition $\alpha_{i1} < \alpha_{i2} < \ldots < \alpha_{im_i}$ must hold. The parameters of the model are estimated using the maximum likelihood estimation

by an E-M algorithm. At the step M of the algorithm a Newton-Raphson iterative scheme is used to solve the non-linear maximum likelihood equation. To score the individuals on the latent variable we can refer to the mean of $z$ defined as:

$$\int z h(z|\mathbf{x}_r) dz$$

where $r = 1, \ldots, n$. These values are normally distributed as given by the previous assumptions.

# 4. Analysis and results

## 4.1 Data collection

Data used in our analysis were collected in various exam sessions of different courses of Computer Science at the University of Bologna. We have grouped the items score together for each argument and normalized it in order to elaborate it by using each model. We have chosen to treat the arguments as "macro items" and to model it as a ordinal items with four answer categories, as described before.

We have considered a sample of 704 students who have written the exam of Computer Science. As for the description of the computer test results, Table 1 shows the percentage and cumulative percentage distributions for each argument.

*Table 1. Percentage and Cumulative percentage distributions*

| | *Category 1* | *Category 2* | *Category 3* | *Category 4* |
|---|---|---|---|---|
| Glossary | 1.70 | 14.63 | 43.18 | 40.48 |
| Prolog1 | 5.97 | 41.62 | 40.48 | 11.93 |
| Prolog2 | 18.89 | 50.57 | 21.73 | 8.81 |
| Prolog | 10.51 | 53.69 | 24.00 | 11.79 |
| Foundat | 10.23 | 52.70 | 33.24 | 3.84 |

We can notice that *Glossary* presents the highest percentages in correspondence of the scores greater or equal to 3. On the contrary, for the *Foundations* and the three arguments concerning *Prolog* (that is *Prolog*, *Prolog1*, *Prolog2*) the most frequent score is 2. It is interesting to notice that the percentage of the students that get high scores for high categories tends to decrease from the first items to the last. That is, it seems to be very important the order of presentation of the items. Perhaps this is a probable explanation of the quite bad performance of the students for the last item. These exploratory results seem to highlight that the items that assess the programming capability and the problem

formalization are more complex to be solved than the items related to the basic knowledge.

## 4.2 Model results

The model estimated has 20 parameters, 3 difficulty parameters for each item and 5 discrimination parameters. The difficulty and discrimination parameter estimates for each item are reported in Table 2.

*Table 2. Parameter difficulty and parameter discrimination estimates*

| | $\alpha_{i1}$ | $\alpha_{i2}$ | $\alpha_{i3}$ | $\beta_i$ |
|---|---|---|---|---|
| Glossary | -4.12 | -1.69 | 0.4 | 0.43 |
| Prolog1 | -2.98 | -0.12 | 2.19 | 0.73 |
| Prolog2 | -2.12 | 1.18 | 3.3 | 1.67 |
| Prolog | -2.46 | 0.68 | 2.32 | 0.94 |
| Foundat | -2.2 | 0.54 | 3.25 | 0.26 |

We can notice that the items that tend to be more difficult are *Prolog1*, *Prolog2* and *Foundat* since their difficulty parameters have very different values among the 3 categories and present the highest absolute values. This indicates that for these items it is very difficult for a student to get a high score and hence to have a good performance. On the other hand, the item that seems to have the lower difficulty is *Glossary* for the opposite reasons listed above. Furthermore we can observe that the item that presents the highest discrimination parameter is *Prolog2* (1.67) followed by *Prolog* (0.941). Conversely, *Foundat* is the item with the lowest discrimination value (0.257). The item *Foundat* needs particular considerations. It is possible that the order of the presentation has played an important role and the statistical design of the experiment necessarily requires to be deeply analyzed.

## 4.3 Student classification

The Italian university mark system fixes in 18/30 the smallest vote to pass an exam, whereas the highest mark is 30/30 *cum laude* (denoted by 30+).
The students can be classified by obtaining a vote according to the different levels of ability. The votes are assigned by fixing two extreme values that indicate, respectively, the lowest level of ability to pass the exam and the highest level of ability. The range between these extreme values is divided into intervals of equal width. A ranked vote is associated to each of them.
On the basis of the indications derived by the previous exams, we decide that the student doesn't pass the exam if ability is smaller than $-0.85$ ($\Phi(z) = 0.20$ where $\Phi$ is the normal distribution function) and we assign the maximum score

(30+) to the students that get a value equal or greater than 1.65 ($1 - \Phi(z) = 0.05$). The remaining votes are given by dividing the range from $-0.85$ to 1.65 into 13 intervals of equal width and by associating them a score graduated from 18 to 30. In this way we can get a rule for student classification that is not influenced by the performance of specific groups of students in different sessions. That is, the vote the student receives is independent from the votes given to the rest of the students so that it is possible that all the students (or none of them) in a session pass the exam.
In Table 3 the frequency distribution of the votes are reported. We can notice that in this session only almost 10% of the students don't pass the exam whereas about the 2% of them get the maximum score. Looking at Figure 1 we can also notice that the distribution of the votes shows a slightly negative skewness, that is, the histogram presents on the left a higher slope than on the right side. This aspect indicates that the students who get a vote ranging from 18 to 23 are the most part.
Table 4 and Figure 4 show the result of a second approach to assign the marks to the examinees. This method is quite similar to the first one, it is different only for the choice of the extremes of the ability intervals. Let $z_{18}$ be the 20th percentile and $z_{30+}$ the 95th of the $z$ distribution, the score is computed by the following linear transformation:

$$Score = 18 + (31 - 18)\frac{z - z_{18}}{z_{30+} - z_{18}} \tag{4}$$

where 31 is the numeric value of 30 *cum laude* (30+).
This last method treats the group of 704 examinees like a statistical population, that is, the value of $z_{18}$ and $z_{30+}$ will play the role of thresholds for the next exams.
In this particular case the results of two methods are quite different because the first one is based on the theoretical normality assumption of the ability and treats the set of students like a sample. Therefore the 20th and the 95th percentile are calculated on the basis of the theoretical normal model.
It is very important to underlie that the differences between the two methodologies seem to decrease if the thresholds $z_{18}$ and $z_{30+}$ are used to determine the marks of another students set.

*Table 3. Classification of students according to their performance (First Method)*

| *Ability* | Mark | $n_i$ | *Ability* | Mark | $n_i$ |
|---|---|---|---|---|---|
| $< -0.85$ | $< 18$ | 76 | $0.50 \vdash 0.69$ | 25 | 42 |
| $-0.85 \vdash -0.66$ | 18 | 40 | $0.69 \vdash 0.88$ | 26 | 30 |
| $-0.66 \vdash -0.47$ | 19 | 57 | $0.88 \vdash 1.07$ | 27 | 35 |
| $-0.47 \vdash -0.27$ | 20 | 101 | $1.07 \vdash 1.27$ | 28 | 30 |
| $-0.27 \vdash -0.08$ | 21 | 77 | $1.27 \vdash 1.46$ | 29 | 12 |
| $-0.08 \vdash 0.11$ | 22 | 76 | $1.46 \vdash 1.65$ | 30 | 6 |
| $0.11 \vdash 0.30$ | 23 | 55 | $\geq 1.65$ | 30+ | 16 |
| $0.30 \vdash 0.50$ | 24 | 51 | | | |

*Table 4. Classification of students according to their performance (Second Method)*

| *Ability* | Mark | $n_i$ | *Ability* | Mark | $n_i$ |
|---|---|---|---|---|---|
| $< -0.59$ | $< 18$ | 140 | $0.41 \vdash 0.55$ | 25 | 40 |
| $-0.59 \vdash -0.45$ | 18 | 47 | $0.55 \vdash 0.70$ | 26 | 33 |
| $-0.45 \vdash -0.30$ | 19 | 61 | $0.70 \vdash 0.84$ | 27 | 20 |
| $-0.30 \vdash -0.16$ | 20 | 59 | $0.84 \vdash 0.98$ | 28 | 29 |
| $-0.16 \vdash -0.02$ | 21 | 62 | $0.98 \vdash 1.12$ | 29 | 18 |
| $-0.02 \vdash 0.13$ | 22 | 61 | $1.12 \vdash 1.27$ | 30 | 19 |
| $0.13 \vdash 0.27$ | 23 | 47 | $\geq 1.27$ | 30+ | 36 |
| $0.27 \vdash 0.41$ | 24 | 32 | | | |

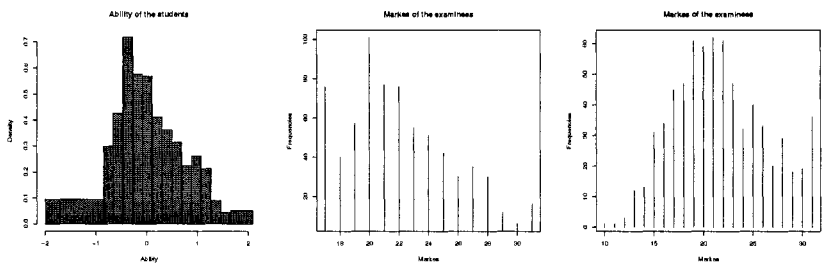


*Figure 1.*

## 5. Conclusions

An IRT model for ordinal observed variables (POM) has been used with the aim to formalize the increasing level of complexity related to the test submitted to the students. Model estimation has highlighted inequalities among the arguments involved in, both in term of difficulty and discrimination. This inequalities have brought out different levels of ability among the students analyzed. It has been also proposed a student classification based on both the

values of the latent variable ability estimated through the model and the judge of the expert, the professor of the subject analyzed.
This is a first analysis of student ability evaluation. Further developments can be considered by improving some aspects of the IRT model, like the goodness of fit problem, and by referring to different data sets in order to consolidate the results. The statistical validation of the evaluation system permits to use this methodology to realize an *assessment in progress*, that is, to give the student a *formative evaluation.*

## References

Gentner, D. and Stevens, A.L. (1983). *Mental Models*. New York: Lawrence Erlbaum Associates.

Lord, F. M. and Novick, M.E. (1968). *Statistical theories of mental test scores*. New York: Addison-Wesley Publishing Co.

Hambleton, R.K. and Swaminathan H. (1985). *Item Response Theory*. Boston: Kluwer - Nijhoff Publishing.

Jöreskog, K. and Moustaki, I. (2001). *Factor Analysis of Ordinal Variables: A Comparison of three Approaches*. Multivariate Behavioral Research, 36, 347-387.

Moustaki, I.(2000). *A Latent Variable Model for Ordinal Variables*.Applied Psychological Measurement, 24, 211–223.

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. Psychometrika Monograph, 17.