Metodološki zvezki, Vol. 2, No. 2, 2005, 209-218

Student Ability Assessment Based on Two IRT Models

Silvia Cagnone[1] and Roberto Ricci[2]

Abstract

The aim of this work is to analyze a part of the data collected in the Computer Science Department during the Informatics exams in the year 2003. Two different Item Response Theory models for ordered polytomous variables are considered in order to get an evaluation of student ability. Ordered polytomous variables are used for a problem solving process that contains a finite number of steps so that the ability of a student can be evaluated on the basis of the step achieved, namely, higher steps achieved are related to higher ability. The models considered are the Partial Credit Model and the Graded Response Model. The choice of these models has been dictated by the fact that although they are defined into different theoretical frameworks, the former belongs to the Rasch family (Masters, 1982) and the latter can be viewed as a Generalized Linear Latent Variable Model (Bartholomew and Knott, 1999), and hence they present different properties, both of them allow to treat ordinal observed variables. The analysis of the real data set through the two approaches allows to highlight their advantages and disadvantages.

1 Introduction

In the last few years, the need for an automated way to assess individual's skills has quickly increased because of the growing request from both private and public structures. Many learning management systems have been developed in order to automatize the learning and assessment process (Gal and Garfiled, 1997; Cagnone *et al.*, 2004). In most of the cases these systems do not allow a quality content evaluation and an efficient evaluation of the student's performance. In the traditional psychometric literature the student's performance is referred indifferently to the terms ability, knowledge, skills, and competence. For this reason from now on we consider these expressions as synonyms.

In the educational systems 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 "not-in-presence" learning programs. Each automatic evaluation system requires the introduction of methodological statistical tools.

In the last decades the problem of assessment has acquired new facets: the self-evaluation, the measuring of both the level of a skill and the effectiveness of a teaching

[1] Department of Statistics, University of Bologna, Italy; cagnone@stat.unibo.it

[2] Department of Statistics, University of Bologna, Italy; rricci@stat.unibo.it

process (Gal and Garfiled, 1997). In the past assessment had just a summative aim, that is the goal of an evaluation was to decide if the examinee was revealing or not a sufficient level of knowledge to pass an exam.

In this new way of understanding the assessment process of a student, evaluation 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 purposes are to identify aspects of performance that need to be improved and to offer corrective suggestions.

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.

The brief description given above makes clear the opportunity to consider both the aspects of the student evaluation. The recent developments in the educational field show that it is necessary to deepen the relation between the pedagogical and the statistical aspects of the assessment.

During his training the student has to pass through different exams. Therefore, it is very important to have at one's disposal assessment methods that are transparent and with solid statistical and pedagogical basis. The weight of this aspect is growing up with the large diffusion of the e-learning products and the computer-automated testing.

Before these 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 of understanding and analyzing a phenomenon. That is, people have different mental models of 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 an examinee performance. We have to conceive it as the exterior expression of a set of latent abilities.

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. A computerized test usually is evaluated by merely making the sum of the obtained score in each question and by translating the final result in one scale representing the human understandable mark. This method does not take into account many aspects strictly inherent to some characteristics of the question like difficulty and discrimination power.

In our particular case we will consider an experimental project developed by Bologna University based on an automatic evaluation system applied in different steps of the educational offer. The principal goal of the project is to assess the students' knowledge in the basic Informatics topics. The data were collected by submitting questionnaires to students who attended preliminary courses of Computer Science.

An ordinal score ranging from 1 to 4 is assigned to each examinee for each item with respect to the solving level achieved. Problems with different steps of complexity have been included in each argument (item). In fact, the problem solving is a process with a finite number of steps. In this way, for every item an ordinal score is assigned to the examinee who completes with success up to a step but fails to complete the subsequent step.

In this work we intend to compare the performances of two Item Response Theory

(IRT) models, the Partial Credit Model (PCM) and the Graded Response Model (GRM), in terms of their advantages and disadvantages in the different applicative steps. Moreover we intend to assess the student ability distribution in the two cases.

2 Model specification

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. One of the principal drawbacks of the classical test theory is that the evaluation of a student's performance is strongly influenced by the sample analyzed. In order to overcome this weakness, at the beginning of the sixieties, a new methodology, called Item Response Theory (IRT) (Lord and Novick, 1968), has been developed. The IRT allows to evaluate the student ability, the question difficulty and the capability of the item to distinguish between examinees with different ability. These properties do not depend on the sample considered.

Since ability is not directly observable and measurable, it is referred to as a 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.

In this paper we apply the PCM and the GRM introduced to treat the case of ordinal observed variables. The aim is to evaluate the possibility of using the two models taking into account the fact that they do not belong to the same theoretical frameworks and hence present different properties. Indeed the PCM belongs to the Rasch family whereas the GRM can be defined within the context of Generalized Linear Latent Variable Models (GLLVM). The GLLVM can be viewed as a general framework within which different kinds of latent variable models are included like the classical factor analysis, when the observed variables are continuous and the IRT models in the case of observed categorical data. The GRM has been formalized within the GLLVM framework by Moustaki (2000) and by Jöreskog and Moustaki (2001).

In the following parts of this section we give a brief description of the two models. The principal features of the PCM and the GRM are detected in the next section where we present an application to a real data set.

2.1 The Graded Response Model

The GRM (Samejima, 1969) is appropriate to use when item answers can be characterized as ordered categorical responses. The GRM can be considered as a generalization of the two parameter model (Birnbaum, 1968) and it belongs to the family of the "indirect" IRT models, in the sense that the computation of the conditional probability for a person responding in a particular category requires two steps.

Each item i is described by one item slope parameter α_i and $j = 1, \ldots, m_i$ between category "threshold" parameters β_{ij}. One goal of fitting the GRM is to determine the location of these thresholds on the latent trait continuum.

The first step consists in the computation of m_i curves for each item according to the following equation

$$P^*_{is}(\theta) = \frac{\exp[\alpha_i(\theta - \beta_{ij})]}{[1 + \exp \alpha_i(\theta - \beta_{ij})]}, \tag{2.1}$$

where θ represents the latent trait (ability).

Each curve in (2.1) describes the probability of a person's item response, denoted by s ($s = j = 1, \ldots, m_i$), falling in or above a given category threshold ($j = 1, \ldots, m_i$), conditional on latent trait level θ.

The GRM is characterized by m_i curves for each item, one curve must be estimated for each between category threshold, that is, m_i parameters β_{ij} and one common α_i slope parameter. The β_{ij} parameters represents the trait level necessary to respond above threshold j with .50 probability.

After the estimation of $P^*_{is}(\theta)$'s, the second step starts. It consists in the computation of the actual category response probabilities for $s = 0, \ldots, m_i$ by the following subtraction

$$P_{is}(\theta) = P^*_{is}(\theta) - P^*_{i(s+1)}(\theta). \tag{2.2}$$

These curves represent the probability that a person answers in a particular category, conditional on the latent trait level θ.

In general, we can say that high values of the slope parameters α_i permit to obtain steep curves given by (2.1) and more narrow and peaked curves given by (2.2). The latter property indicates that the response categories differentiate among latent trait level fairly well.

The β_{ij}'s determine the location of the curves (2.1) and where each of the curves (2.2) for the middle answer options peaks, i.e. the curves (2.2) peak in the middle of two subsequent threshold parameters.

As mentioned before, one of the appealing aspect of the GRM is that it can be viewed as a GLLVM (Bartholomew and Knott, 1999). This represents a general framework in which different statistical methods, including the IRT, are conveyed. A substantial difference between the GLLVM approach and the IRT approach is that in the former the latent trait is treated as a multidimensional random variable so that more abilities underlying the learning process can be investigated.

More in detail, the aim of the GLLVM is to describe the relationship between p manifest variables x's and $q < p$ latent variables θ's in the following way

$$f(\mathbf{x}) = \int g(\mathbf{x} \mid \boldsymbol{\theta}) h(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}, \tag{2.3}$$

where $h(\boldsymbol{\theta})$ is assumed to be a standard multivariate normal distribution and $g(\mathbf{x} \mid \boldsymbol{\theta})$ is assumed to be a member of the exponential family. Moreover the conditional independence of the observed variables given the latent variables is assumed so that

$$g(\mathbf{x} \mid \boldsymbol{\theta}) = \prod_{i=1}^{p} g(x_i \mid \boldsymbol{\theta}). \tag{2.4}$$

With reference to the GRM described above, $g(x_i \mid \boldsymbol{\theta})$ is a multinomial probability function (Moustaki, 2000)

$$g(x_i \mid \boldsymbol{\theta}) = \prod_{s=0}^{m_i} [P^*_{is}(\boldsymbol{\theta}) - P^*_{i(s+1)}(\boldsymbol{\theta})]^{x_{is}}, \tag{2.5}$$

where $x_{is} = 1$ if the response falls in category s of the item i and $x_{is} = 0$ otherwise. The relation between the observed and the latent variables is expressed in (2.1) in the unidimensional case.

2.2 The Partial Credit Model

The PCM (Masters, 1982) was originally developed for analyzing items for which it is important to assign partial credit for completing several steps in the solution process. That is, the PCM is naturally thought for describing item answers where it is important to assess the response not according to the conceptual structure right/wrong, but where partially correct answers are permitted.

The PCM is a “direct” IRT model, that is, the probability of answering in a category is given directly as an exponential expression divided by the sum of the exponentials. Furthermore the PCM can be considered as an extension of the Rasch model and it conserves its main features such as separability of person and item parameters.

Assume that item i is scored by $s = 0, \ldots, m_i$. For $s = j$ the category response curves for the PCM can be written as

$$P_{is}(\theta) = \frac{\exp\left[\sum_{j=1}^{s}(\theta - \delta_{ij})\right]}{\sum_{r=1}^{m_i+1}\left[\exp\sum_{j=1}^{r}(\theta - \delta_{ij})\right]}. \tag{2.6}$$

A δ_{ij} ($j = 1, \ldots, m_i$) term can be directly interpreted as the item step difficulty associated with a category score of j, that is, δ_{ij} parameters can be considered as step “difficulties” associated with the switch from one category to the next. There are m_i step difficulties for an item with $m_i + 1$ response category.

In the PCM, like in all the polytomous Rasch models, the δ_{ij} parameters do not represent a point on the latent trait scale at which a student has a .50 probability of responding above a category threshold, as the β_{ij} parameters do in the GRM, but they point out the relative difficulty of each step.

3 Analysis and results

3.1 Data description

The data used in our analysis were collected in various exam sessions of Bologna University in the courses of basic Computer Science. We have considered a sample of 704 students who have written the exam of Computer Science. In particular, the test sessions have been organized by using a database that contains 5 different arguments: *Glossary* has to do with the meaning of some words or the functions of some objects of the computer world, *Foundations* regards the basic knowledge on calculability, algorithms complexity, computer architecture, compiler, and programming languages, and *Prolog* items are concerned with several aspects of the programming reasoning.

The problem solving process contains a finite number of steps so that the ability of a student can be evaluated on the basis of the step achieved, namely, higher steps achieved are related to higher ability. 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).

As for the description of the computer test results, Table 1 shows the percentage and cumulative percentage distributions of the answers to each argument.

Table 1: Percentage and Cumulative percentage distributions.

	Category 1		Category 2		Category 3		Category 4	
	%	cum %	%	cum %	%	cum %	%	cum %
Glossary	1.70	1.70	14.63	16.34	43.18	59.52	40.48	100
Prolog1	5.97	5.97	41.62	47.59	40.48	88.07	11.93	100
Prolog2	18.89	18.89	50.57	69.46	21.73	91.19	8.81	100
Prolog0	10.51	10.51	53.69	64.20	24.00	88.21	11.79	100
Foundations	10.23	10.23	52.70	62.93	33.24	96.16	3.84	100

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 Prolog0, 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 ones. That is, it seems to be very important the order of presentation of the items. 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.

3.2 Model results

The following table shows the results concerning the parameters of the models (2.2) and (2.6) estimated through the Marginal Maximum Likelihood method by using the software MULTILOG 7.0.3.

Table 2: Parameter Estimation.

	GRM				PCM		
	α	β_{i1}	β_{i2}	β_{i3}	δ_{i1}	δ_{i2}	δ_{i3}
Glossary	0.43	-9.56	-3.89	0.91	-2.64	-1.25	0.12
Foundations	1.61	-1.31	0.71	1.99	-1.11	0.99	1.11
Prolog1	1.00	-2.50	0.69	2.34	-1.87	0.82	1.02
Prolog0	0.73	-4.07	-0.19	2.97	-2.19	-0.02	1.43
Prolog2	0.30	-7.32	1.79	10.70	-1.92	0.48	2.54

As we pointed out before, the δ's parameters in polytomous Rasch Models, such as the PCM, do not represent a point on the latent trait scale at which an examinee has a .50

probability of responding above a category threshold, as the β's parameters do in the GRM. However the analysis of the parameter values of the two models allows to get very similar orders of the item difficulty. Indeed, if we take into account the width of the δ's intervals of the PCM and the β's values of the GRM, we can obtain a very similar increasing difficulty ranking. It is possible to observe that both the models indicate the same questions at the extremities of the difficulty range: Glossary at the bottom and Prolog2 at the top. But the GRM allows a further analysis about the discrimination power of each question. Moreover it is remarkable that Glossary and Prolog2 have a quite low value of the α parameter. This aspect suggests that it could be of some interest to think about changes of the questions at the extremities of the increasing difficulty ranking.

Since the principal aim of the experimental project of Bologna University is to improve an automatic tool for the evaluation, it is important to analyze the distributions of the estimated abilities obtained by the PCM and the GRM. Figure 1 depicts the estimated abilities according to the two models and it is evident that the results are quite similar. More in detail in both cases the ability distribution is rather symmetric, highlighting more capability to distinguish among students that present high ability. As found before (Table 2), this is mainly due to the fact that the easier items do not allow to catch the differences among low values of ability scale.

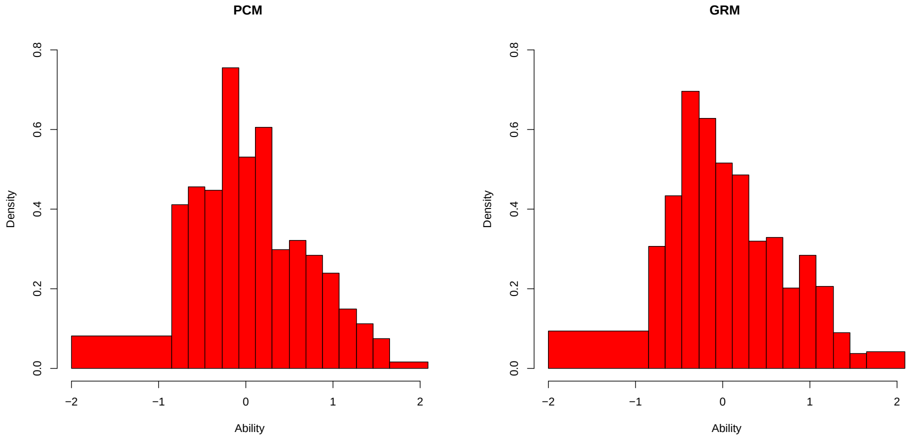

Figure 1: Ability distributions.

A further evidence of the similarity between the two distributions is shown in Figure 2 in which each point represents the individual's ability estimated by using the two models.

This result brings into prominence that the choice of a model against the other is mainly dictated by the research goals (preliminary item calibration, ability estimation, and so forth) rather than the performance differences.

As for the evaluation of the goodness of fit, in Table 3 the observed and expected proportions in each category of all the items by using both the models are reported. The evident proximity of the parameter values of the expected proportions of the PCM and GRM seems to indicate a good fit of the both of them to the data. Nevertheless, for the IRT models the goodness of fit is still an open issue. In literature some theoretical solutions (Jöreskog and Moustaki, 2001; Mignani and Cagnone, 2004) have been pointed out with reference to GLLVM for ordinal data.

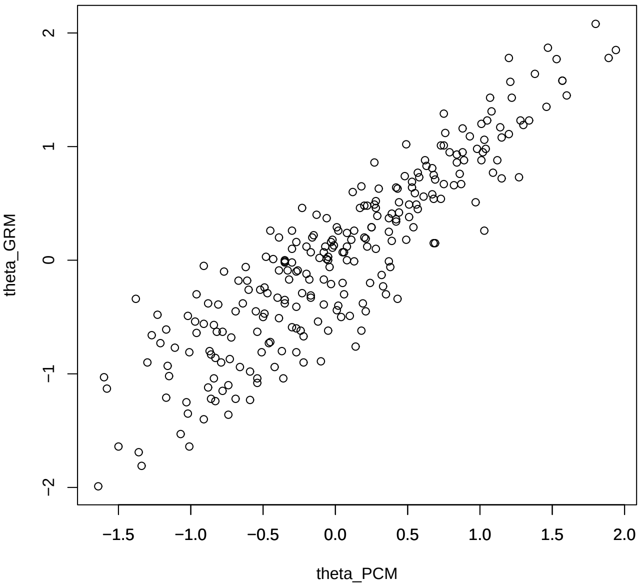

Figure 2: Scatter plot of the estimated abilities (GRM versus PCM).

Anyway further investigation is needed in terms of the performance comparison of the two models.

4 Concluding remarks

The analysis allows to evaluate the student ability concerning Computer Science problems by using a computer test delivery. Parameter estimations obtained by using both the models feature inequalities among the arguments involved in. These differences concern the item difficulty and, if the GRM is used, the discrimination parameters.

Although the results obtained evidence quite similar behaviour of the PCM and the GRM, some differences are remarkable. They are quite little in practice, but they have important methodological and conceptual meanings. First, the presence of a discrimination parameter in the GRM could be very useful in the preliminary calibration phase relevant for the questionnaire design. In fact the application of the GRM could represent an efficient tool to select appropriate items. Second, when well calibrated items are chosen for the questionnaire, it could be recommended to estimate the individual's ability through the PCM since, as mentioned before, this model belongs to the Rasch family. As known in literature, the Rasch models have very interesting statistical and computational proprieties.

However, the possibility to locate the GRM in the context of the GLLVM allows to use the theoretical properties of this approach. Although not applied in this paper, the GLLVM permit to estimate more than one ability and to investigate the potential correlation between them. Furthermore, as said before, some results concerning the goodness of fit evaluation of models for ordinal data have been obtained in literature and several

Table 3: Observed and expected proportions according to PCM and GRM.

		Cat. 1	Cat. 2	Cat. 3	Cat. 4
Glossary	Obs.	.0170	.1477	.4290	.4069
PCM	Exp.	.0172	.1476	.4271	.4081
GRM	Exp.	.0171	.1472	.4288	.4069
Prolog0	Obs.	.0597	.4134	.4063	.1207
PCM	Exp.	.0597	.4113	.4069	.1221
GRM	Exp.	.5980	.4098	.4097	.1207
Foundations	Obs.	.1875	.5057	.2159	.0909
PCM	Exp.	.1873	.5039	.2179	.0909
GRM	Exp.	.1865	.5011	.2225	.0899
Prolog1	Obs.	.1051	.5355	.2401	.1193
PCM	Exp.	.1052	.5337	.2420	.1192
GRM	Exp.	.1056	.5326	.2419	.1199
Prolog2	Obs.	.1023	.5270	.3310	.0398
PCM	Exp.	.1039	.5254	.3342	.0395
GRM	Exp.	.1025	.5263	.3314	.0398

studies are in progress. This problem is still troublesome in the IRT and at the moment it is an open research question.

Acknowledgement

The authors thank prof. Stefania Mignani for her useful suggestions and remarks.

References

[1] Bartholomew, D. and Knott, M. (1999): *Latent Variable Models and Factor Analysis*. London: Kendall's Library of statistics.

[2] Cagnone, S., Mignani, S, Ricci, R., Casadei, G., and Riccucci, S. (2004). Computer-Automated testing: an evaluation of undergraduate student performance. *Proceedings of Technology Enhanced Learning*, 59-68, Berlin-Heidelberg: Springer-Verlag.

[3] Gal, I. and Garfiled, G.B. (Eds.) (1997): *The Assessment Challenge in the Statistical Education*. Amsterdam: IOS Press.

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

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

[6] Jöreskog, K. and Moustaki, I. (2001): Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research*, **36**, 347-387.

[7] Lord, F.M. and Novick, M.E. (1968): *Statistical Theories of Mental Test Scores*. New York: Addison-Wesley Publishing Co.

[8] Masters, G.N. (1982): A Rasch Model for Partial Credit Scoring. *Psychometrika*, **47**, 149-174.

[9] Mignani, S., Cagnone, S., Casadei, G., and Carbonaro, A. (2005): An Item Response Theory Model for student ability evaluation using computer-automated test results. In Vichi, Monari, Mignani, Montanari (Eds.): *New Developments in Classification, Data Analysis, and Knowledge Organisation*, 321-329. Berlin-Heidelberg: Springer-Verlag.

[10] Mignani, S. and Cagnone, S. (2004): A comparison among different solutions for assessing the goodness of fit of a generalized linear latent variable model for ordinal data. *Statistica Applicata*, **16**,1-19.

[11] Moustaki, I. (2000): A Latent Variable Model for ordinal variables. *Applied Psychological Measurement*, **24**, 211-223.

[12] Parshall C., Spray J., Kalohn J., and Davey T. (2002): *Consideration in Computer-Based Testing*. New York: Springer-Verlag.

[13] Samejima, F. (1969): Estimation of ability using a response pattern of graded scores. *Psychometrika Monograph*, **17**.

[14] Van der Linden, W.J. and Hambleton R.K. (Eds.) (1997): *Handbook of Modern Item Response Theory*. New York: Springer-Verlag.