Quaderni di Statistica
Vol. 11, 2009

A proposal for classifying students' maths performances in lower secondary school

Mariagiulia Matteucci Stefania Mignani
Statistics Department, University of Bologna
E-mail: m.matteucci@unibo.it, stefania.mignani@unibo.it

Summary: In this work, a combined use of different methods is proposed in order to evaluate mathematics students' achievement within the Italian lower secondary school. Despite the increasing need for an objective assessment of student performance, a univocal approach still does not exist. By using data from the INVALSI standardized test administered in 2008 to Italian lower secondary school students, a joint approach of test score, latent trait and latent class analysis is proposed which correctly highlights both item features and student differences. The methods show an effective capability of differentiating the examinees into three performance groups on the basis of the response patterns.

Keywords: latent variable models, mathematics assessment, student evaluation.

1. Introduction

In the Italian educational system, the requirement for a rigorous framework to assess formative outcomes is increasingly important. In this context, the evaluation of learning achievement is one of the most relevant aspects required in order to understand strengths and isolate weaknesses. The evaluation of student learning and competence, *i.e.* the capability of combining knowledge and ability elements in a specific context effectively, can be typically verified by administering a questionnaire whose items are related to target skills. This process involves a number of methodological tools as regards data collection, test design compatibility and statistical analysis application.

The National Evaluation Institute for the School System (INVALSI) regularly carries out large-scale national evaluations to identify patterns and trends in the achieved competence of pupils attending both primary and lower secondary schools. In the scholastic year 2007-2008, in addition to the traditional final written and oral examination, a standardized test for evaluating competency in mathematics, reading comprehension and Italian grammar was administered to all Italian students in the lower secondary school for the first time.

The present study has both an empirical and a methodological aim. The empirical aim is to find an objective criterium for judging pupils. This involves the assignment of a score indicating the performance level of each student, and successively the choice of one or more thresholds for classifying the outcomes. The methodological aim is to discuss the techniques of analysis which satisfy the empirical research questions. To this end we compare three different approaches: the first based on the summed test score in the context of classical test theory and the remaining two developed in the latent variable model framework. In particular, item response theory models and latent class analysis are considered (Bartholomew and Knott, 1999). Starting from the preliminary work of Matteucci, Mignani, and Ricci (2009), we suggest a joint use of the three approaches for identifying a classification rule which is both reliable and flexible in order to discern relevant differences in learning achievement.

The paper is organized in the following way. In Section 2, the INVALSI test concerning mathematics is presented and some descriptive indicators on item responses are reported. Section 3 describes the essential methodological facets of the analyses and highlights the results obtained within each approach. Finally, Section 4 shows our innovative proposal for the classification of respondents.

2. Data and descriptives

The data used in our study come from the INVALSI test taken at the end of the scholastic year 2007-2008. The test was designed in such a way that questions, administering conditions, scoring procedures and interpretations were consistent for all the Italian schools. Approximately

560,000 students in 5,923 different schools received the questionnaire. In the current work, we consider a random sample obtained through a three-way stratification sampling (school, class, student) with reference to five Italian geographic areas: North-East, North-West, Center, South and Islands. In order to work with a homogeneous sample of students under equivalent testing conditions, disabled students were excluded because they had more time to complete the test. The sample size is of 4,881 test-takers. The sample is used mainly to reduce the computational efforts due to the large size of the population and also to analyze the achievement results more in detail. An empirical and feasible classification rule which can be extended to the whole population of students is developed.

The test includes two sections: the first one assessing reading comprehension and Italian grammar competence, and the second one evaluating mathematical skills. In this paper, we focus on the mathematics test which consists of twenty-two items concerning common topics taught in the lower secondary school. These topics were consistent with those used in large-scale international assessment programs such as PISA and TIMSS and included some national specificities.

In detail, the item domains deal with number (N), geometry (G), functions and relationships (FR), measurement and data (MD). The last domain includes statistical and probability questions. Several types of items are designed: multiple-choice with four alternatives with only one correct answer, true-false and open-ended items that ask students to give both a numeric answer and the adopted procedure. For binary and multiple choice items a score equal to one is assigned if the answer is correct and zero otherwise, while for open-ended items a graded score is assigned (2 = both correct answer and procedure, 1 = incorrect answer but correct procedure, 0 = both incorrect answer and procedure). The correction grid assumes that the same distance between the graded scores holds. This assumption is quite strong and can be overcome by using a probabilistic model which estimates the item properties after the response patterns have been observed (see Section 3.2).

A preliminary classical analysis of the item responses is carried out. The Cronbach Alpha is equal to 0.79 indicating a high level of test reliability. The relative frequencies of correct responses for each item are

reported in Table 1. The test results concerning the whole population of candidates can be found in Matteucci and Mignani (2009).

Table 1. Relative frequencies of correct response for each item

Item	Prop. correct	Item	Prop. correct
1N	0.780	3MD	0.551
5N	0.180	7MD	0.710
8N	0.386	19aMD	0.632
10N	0.082*	19bMD	0.054*
	0.279		0.173
14N	0.507	21MD	0.089*
15N	0.431		0.405
17N	0.731		
		12FR	0.734
2G	0.661	18FR	0.823
4G	0.719	20FR	0.739
6G	0.717		
9G	0.590		
11G	0.274		
13G	0.671		
16G	0.847		

* denotes a partially correct response.

As we can expect from a well calibrated test, the frequencies are quite different, ranging from 0.173 to 0.847. Item 16 on geometry obtains the highest number of correct answers while item 19b in the measurement and data domain shows the worst performance. This particular item requires the computation of the arithmetic mean in a frequency distribution. It should be noted that, within each domain, items show different performances, except for the three function and relationship items that show similar response frequencies. At first sight, these descriptives suggest the need for further analysis which takes into account the item properties.

3. Student evaluation: methods and analyses

To date, various methods have been developed and introduced in order to assess students' performances. One of the more practical and intuitive ways of evaluating test results is by assigning a score based on the number of correct responses or on the steps taken to reach a solution. Within this approach, a useful indicator is the total test score (TTS), which assigns a raw score to each individual. Another possible solution for student evaluation can be found in latent variable models (Bartholomew and Knott, 1999), which assume the existence of a single or multiple unobserved variables, called *abilities*, which account for the covariation among the item responses. Depending on the assumption on the latent variables, item response theory (IRT) models or latent class analysis (LCA) can be considered. The main difference between IRT (Lord and Novick, 1968) and LCA (Lazarsfeld and Henry, 1968) is in the assumption of ability distribution, which is continuous in the former while concentrated in a small number of discrete points in the latter. In practice, IRT allows for the individual scoring based on a particular model, while LCA is used in order to assign students to a particular group.

3.1. Total test score

In the common practice of student evaluation, one of the most widely employed measures is undoubtedly the total test score. The idea of using TTS in educational measurement was developed within the classical test theory (CTT) founded on the concept of true score (Novick, 1966; Lord and Novick, 1968). The true score is defined as the expected value of the observed scores over an infinite run of independent repeated observations and the fundamental equation of CTT expresses the observed score as the sum of the true score and a random error.

Given a set of k binary items in a test, the TTS for person i, with $i = 1, ..., n$, is simply the sum of correct responses, as follows

$$TTS_i = \sum_{j=1}^{k} Y_{ij}, \tag{1}$$

where Y_{ij} is the response variable of examinee *i* to item *j*, with $j = 1, ..., k$ items, taking value 1 for a correct response and value 0 for an incorrect response. More generally, the total test score is the sum of the individual item scores, depending on how the responses have been scored.

In our case study, the mathematics items have been scored as 0, 1 or 2 depending on the correction grid. Therefore, the variable Y_{ij} in Equation (1) may take integers in the range [0;2]. Because the test consists of twenty-two items and three of them are graded, the TTS has a minimum value of 0 and a maximum of 25, as shown in Figure 1.

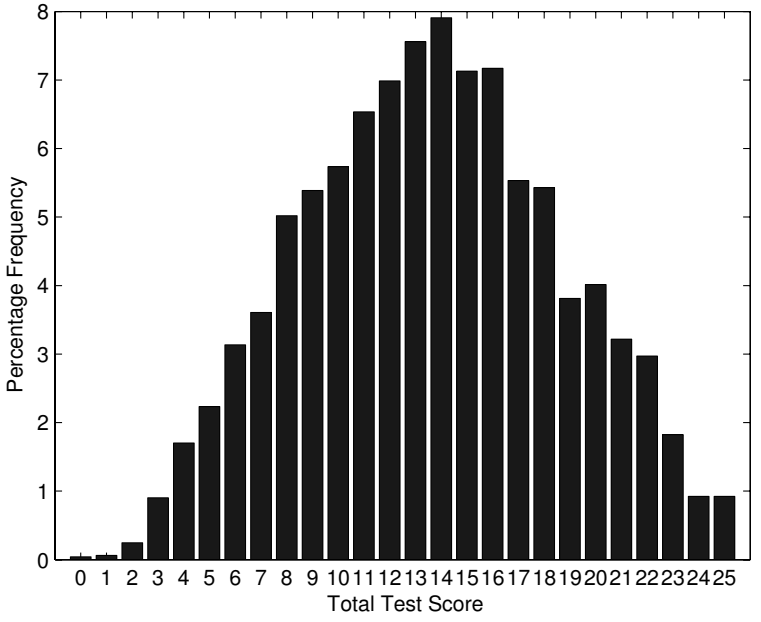

Figure 1. Percentage distribution of total test score

The mean value reported for TTS is 13.62 while the median is equal to 14, with a standard deviation of 4.99. In particular, only two students

answered all items incorrectly, while forty-five students completed the entire test successfully. By using the total test score, only a raw score is assigned to each examinee, which does not take into account the item characteristics, such as the item difficulty, but only the item number and the score given in the correction grid. In order to include the item features in the assignment of individual scores, we can resort to item response theory models both for the item analysis and the student evaluation.

3.2. IRT-based scoring

In order to take into account the item properties in the assessment of performances, an item response theory model is considered. IRT is sometimes assimilated to latent trait analysis (LTA), denoting that categorical responses are treated by using continuous latent variables. IRT models allow the simultaneous estimation of the item characteristics (*e.g.* difficulty and discrimination) and individual abilities. In particular, because the test consists of binary and graded items, a general model for ordered responses is needed. Under the assumption of unidimensionality, *i.e.* a single latent variable θ underlying the response process, the graded response model (GRM) due to Samejima (1969) properly fits the data.

According to the GRM, the probability of a response in category s, with $s = 1, ..., m_j$ response categories for item j and $j = 1, ..., k$ items, is given by

$$P(Y_j = s|\theta) = P^*_{js}(\theta) - P^*_{j(s+1)}(\theta), \tag{2}$$

where $P^*_{js}(\theta)$ is called *operating characteristic curve* (Embretson and Reise, 2000) and describes the probability of a response in category s or higher as

$$P^*_{js} = P(Y_j \geq s|\theta) = \frac{\exp[\alpha_j(\theta - \beta_{j(s-1)})]}{1 + \exp[\alpha_j(\theta - \beta_{js})]}. \tag{3}$$

In Equation (3), α_j is the discrimination parameter, which gives a measure of the capability of the item to differentiate between the examinees, and β_j is the threshold (or difficulty) parameter representing the difficulty

of the item steps (*e.g.* $\beta_{j(s-1)}$ is the level of ability required to have 50% probability to score *s* or higher). For each item, one discrimination parameter and $m_j - 1$ threshold parameters are estimated. The GRM is a "two-step" model, because the computation of the *category response curves* in (2) requires first the computation of the $m_j - 1$ operating characteristic curves in (3) and then, by difference, the response probabilities according to (2). The model is widely employed when dealing with ordered polytomous items, and is able to handle binary items as a special case. In fact, the GRM has been developed as a generalization of the two-parameter logistic (2PL) model (Birnbaum, 1968). The model is estimated, under the assumption of local independence, by using marginal maximum likelihood via EM algorithm, choosing a standard normal distribution as a prior for θ.

Table 2 shows the item parameter estimates according to the GRM for the mathematics test.

Table 2. Item parameter estimates according to the GRM (standard errors in brackets)

Item	$\hat{\alpha}$	$\hat{\beta}$	Item	$\hat{\alpha}$	$\hat{\beta}$
1N	0.796 (0.05)	-1.792 (0.10)	3MD	0.718 (0.05)	-0.319 (0.05)
5N	0.769 (0.05)	2.198 (0.12)	7MD	1.446 (0.06)	-0.857 (0.04)
8N	1.083 (0.05)	0.523 (0.04)	19aMD	1.289 (0.06)	-0.559 (0.03)
10N	1.476 (0.06)	0.535 (0.03)	19bMD	1.403 (0.06)	1.161 (0.04)
		0.889 (0.13)			1.478 (0.35)
14N	0.714 (0.04)	-0.045 (0.05)	21MD	1.298 (0.05)	0.014 (0.02)
15N	0.870 (0.04)	0.365 (0.04)			0.382 (0.06)
17N	1.355 (0.06)	-0.989 (0.04)			
			12FR	0.867 (0.05)	-1.353 (0.07)
2G	0.848 (0.04)	-0.913 (0.05)	18FR	1.032 (0.06)	-1.783 (0.08)
4G	0.713 (0.04)	-1.461 (0.09)	20FR	0.666 (0.04)	-1.710 (0.11)
6G	1.572 (0.07)	-0.851 (0.03)			
9G	1.041 (0.05)	-0.433 (0.04)			
11G	0.992 (0.05)	1.172 (0.06)			
13G	0.982 (0.05)	-0.872 (0.05)			
16G	0.795 (0.05)	-2.403 (0.14)			

For items recoded as binary, one discrimination parameter and one difficulty parameter are estimated, while for the graded items (10, 19b, 21) one discrimination parameter and two thresholds are estimated. It can be seen that the discrimination parameter estimates are all positive and moderately high (above 0.7), denoting a proper capability of the items to differentiate between the examinees. In particular, for the open constructed-response items, the discrimination estimates are all observed to be rather high, suggesting that this item type is particularly adept to catch differences in ability. As regards the difficulty parameters β, the estimates cover a wide range of values in the ability scale, from -2.403 (item 16) to 2.198 (item 5), denoting items with different levels of difficulty. In particular, item 16 on geometry is estimated the easiest while item 5 on numeracy is the most difficult.

In order to characterize the examinees' abilities, different scoring methods may be applied adopting a frequentist or a Bayesian approach (for a review see Baker and Kim, 2004). One of the most common choices in practice is to adopt the Bayes expected a posteriori (EAP) estimator, as follows

$$E(\theta|\mathbf{y_i}) = \int_{\Theta} \theta P(\theta|\mathbf{y_i}) d\theta, \tag{4}$$

where $\mathbf{y_i}$ is the vector of item responses for person *i*. The EAP estimator finds the mean of the posterior distribution of ability, given the observed response pattern, by using Gaussian quadrature.

The popularity of this estimator is motivated by the fact that, as discussed in Bock and Mislevy (1982), the EAP estimator has minimum square error over the population of ability and it is the most accurate on average. Moreover, the estimator does not require iterative procedures and it is defined for perfect response patterns. EAP ability scores have been estimated for each respondent and the percentage distribution is given in Figure 2. Typically, due to the standard normal assumption for the ability prior distribution, the estimated scores are included in the range [-3;3] but values in the extremities are estimated with less precision. For the respondents to the mathematics test, the minimum estimated score is -2.77 while the maximum is 2.31. The mean value is 0.00 while the me-

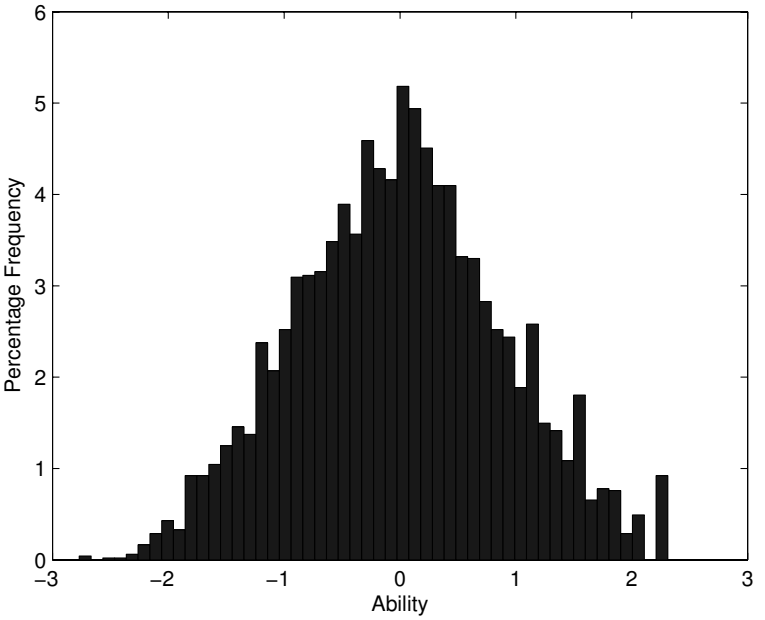

Figure 2. Percentage distribution of EAP scores

dian is 0.01 with a standard deviation of 0.90. The 25th percentile is -0.63 while the 75th is 0.60 so that the distribution is balanced.

Unlike the total test score, which is simply an aggregation of the item responses, IRT scores are based on the item characteristics which means that respondents with the same summed score but different response pattern may have different estimated ability (van der Linden and Hambleton, 1997).

3.3. Latent class analysis

Within the framework of latent variable models, latent class analysis (Lazarsfeld and Henry, 1968), has been developed in order to find latent groups in which individuals could be allocated. Unlike IRT, which assumes continuous latent variables, latent class analysis is based on the ex-

istence of categorical latent variables, consisting of discrete points. Analogously to Section 3.2, a single latent ability is assumed, but taking discrete values $\theta = c$, with $c = 1, ..., C$ latent classes. LCA is founded on the assumption of conditionally independence, which implies that responses are stochastically independent, given the class membership.

Given the item response variable Y_{ij} taking values $s = 1, ..., m_j$, the focus of LCA is on the conditional response probabilities to item *j* given membership in class *c*, $\pi_{sc} = P(Y_j = s|\theta = c)$, and on the prior probabilities of belonging to class *c*, $\eta_c = P(\theta = c)$. Therefore, the basic unrestricted model for LCA is

$$P(Y_i = s) = \sum_{c=1}^{C} P(\theta = c) \prod_{j=1}^{k} P(Y_{ij} = s|\theta). \tag{5}$$

The estimation of latent class models does not require assumptions on the prior distribution of the latent variable and can be conducted by using maximum likelihood via EM algorithm. In practice, only the number of classes should be specified. Usually, the rule is to start from two classes and gradually increase the number of classes until is possible to find the point at which the model fits the data better.

In our case study, sparse data (response patterns with low observed frequencies) represent a serious obstacle in order to check the model fit. In fact, on a sample of 4,881 students, we have 4,363 different response patterns. Solutions of up to 7 classes were estimated, but global fit indexes as Pearson χ^2 or Likelihood-ratio could not be used. An alternative was to choose the model associated to the minimum value for information indexes such as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC). Unfortunately, increasing the number of classes decreased both information criteria, although not considerably. Therefore, we have decided to take into account a three-class solution, which guarantees an high interpretability and implies a minimum decrease in information indexes. Table 3 reports the estimated conditional response probabilities for each item and response category. For items recoded as binary, only the probability of a correct response is reported, conditionally to each latent class, while for graded items, the first condi-

tional probability refers to a partially correct response and the second one to a correct response.

Table 3. Estimated conditional response probabilities (standard errors in brackets)

Item	$\hat{\pi}_{s1}$	$\hat{\pi}_{s2}$	$\hat{\pi}_{s3}$	Item	$\hat{\pi}_{s1}$	$\hat{\pi}_{s2}$	$\hat{\pi}_{s3}$
1N	0.93 (0.01)	0.81 (0.01)	0.60 (0.02)	3MD	0.81 (0.02)	0.55 (0.01)	0.35 (0.02)
5N	0.42 (0.02)	0.13 (0.01)	0.10 (0.01)	7MD	0.97 (0.01)	0.78 (0.01)	0.37 (0.02)
8N	0.77 (0.02)	0.36 (0.02)	0.15 (0.01)	19aMD	0.92 (0.01)	0.70 (0.02)	0.28 (0.02)
10N	0.11 (0.01)	0.10 (0.01)	0.02 (0.01)	19bMD	0.08 (0.01)	0.06 (0.01)	0.02 (0.01)
	0.67 (0.02)	0.26 (0.01)	0.03 (0.01)		0.52 (0.03)	0.12 (0.01)	0.01 (0.00)
14N	0.79 (0.02)	0.49 (0.01)	0.32 (0.02)	21MD	0.06 (0.01)	0.12 (0.01)	0.06 (0.01)
15N	0.74 (0.02)	0.43 (0.01)	0.20 (0.01)		0.79 (0.02)	0.41 (0.02)	0.09 (0.01)
17N	0.97 (0.01)	0.80 (0.01)	0.41 (0.02)				
				12FR	0.93 (0.01)	0.77 (0.01)	0.51 (0.02)
2G	0.88 (0.01)	0.70 (0.01)	0.42 (0.02)	18FR	0.98 (0.01)	0.87 (0.01)	0.62 (0.02)
4G	0.88 (0.01)	0.76 (0.01)	0.52 (0.02)	20FR	0.90 (0.01)	0.76 (0.01)	0.57 (0.02)
6G	0.97 (0.01)	0.81 (0.01)	0.34 (0.02)				
9G	0.90 (0.01)	0.62 (0.02)	0.30 (0.02)				
11G	0.57 (0.02)	0.26 (0.01)	0.08 (0.01)				
13G	0.91 (0.01)	0.71 (0.01)	0.40 (0.02)				
16G	0.96 (0.01)	0.87 (0.01)	0.71 (0.02)				

As shown in Table 3, the three latent groups highlight different relevant behaviors in the test responses. In particular, the first class denotes the students with the best performances, *i.e.* with the highest probabilities of giving a correct answer to all the items. In contrast, the third class represents the students with the weakest results while the second group shows an intermediate level of performance. The prior probabilities are estimated as $\eta_1 = 0.20$, $\eta_2 = 0.54$ and $\eta_3 = 0.26$. Once the conditional probabilities have been estimated, individuals can be classified through the computation of the posterior probabilities of belonging to a specific class, given the response pattern $P(\theta = c|Y_1, ..., Y_k)$.

The results show that as well as being used for the allocation of individuals, LCA is also a power instrument for investigating the test structure in terms of item characteristics. Unlike IRT-based scoring which provides each individual with a score on the real line, LCA has been used to assign each test-taker to a latent group. LCA assumes a lower level of detail in the performance of the candidate but is a more straightforward classification instrument.

4. A proposal of student classification

In order to define an empirical criterion for classifying students, the scoring results of the three approaches in Section 3 have been analysed simultaneously. Figure 3 displays the scatter plot of estimated ability and total score for each student.

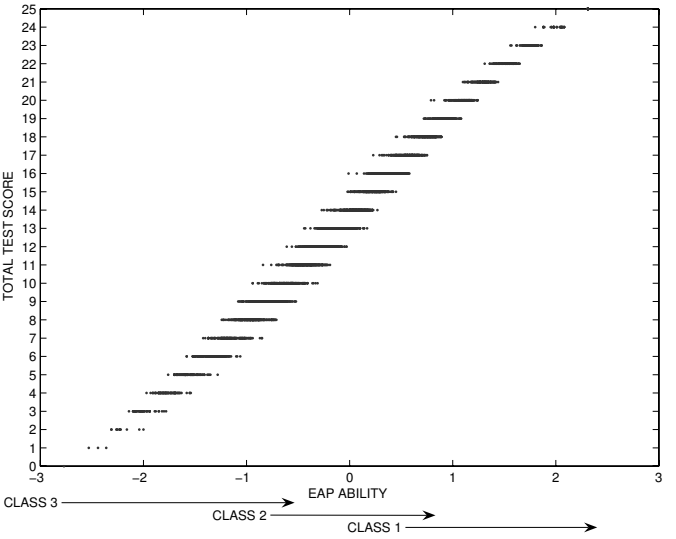

Figure 3. Scatter plot of EAP score vs. total test score

As expected, ability and total score have a similar behaviour, nevertheless students with equal summed score can have different estimated ability, depending on the particular response pattern. As is well-known, the information obtained by the total test score is partial while IRT scoring allows a more detailed graduation in the different performances of respondents. As regards the estimated latent classes, the arrows indicate the range of ability values estimated within each class. As can be seen, the overlapping between the arrows is very small, denoting a rather precise classification. Considering TTS, EAP scores and latent classes jointly, we can summarize the results in Table 4.

It can be noted that the results are overlapped partially. In particular,

Table 4. Score estimates within each class

Class	TTS range	EAP score range
1	17 - 25	0.61 - 2.31
2	9 - 19	-0.74 - 0.93
3	0 - 11	-2.77 - -0.45

with the intention to assign students to the most favourable latent class, we have found out that 443 students with a TTS between 17 and 19 belong to class 2 instead of class 1, while 444 students with a score between 9 and 11 belong to class 3 instead of class 2. As regards EAP score, an overlapped set of 219 students with score between 0.61 and 0.93 is in class 2 rather than class 1 while 215 students with a score between -0.74 and -0.45 are in class 3 instead of class 2.

In consequence of these results, we propose different steps for classifying students. First, we suggest applying latent class in order to identify a wide-ranging classification. In such a way students with similar performances are grouped together, as in our case study, where the three groups represent pupils with the worst, medium and best outcomes. Second, estimating the EAP scores we can differentiate more in detail the behaviour of students in the same group and performances can be graduated assigning a proper score to all respondents. Considering the results jointly and analyzing the students overlapping, respondents should be assigned to the most favourable class.

To sum up, our solution seems to be flexible for taking into account several aims of learning assessment and obtaining a rigorous criterion with no computational efforts. This approach may be extended to other standardized tests (*e.g.* the reading comprehension test). Nevertheless, further insights are needed in order to verify this strategy both from a methodological and an empirical point of view. In particular, the strategies of student classification always depend on the goals of the examiner. For this reason, our statistical analysis should be compared to educational expert opinion. Finally, an interesting research question is related to the effect of the item type and the item domain on the student classification.

Acknowledgements: The authors gratefully acknowledge the National Evaluation Institute for the School System (INVALSI) for data availability.

References

Baker F. B., Kim S. H. (2004), *Item Response Theory: Parameter Estimation Techniques*, Marcel Dekker, New York.

Bartholomew D. J., Knott M. (1999), *Latent Variable Models and Factor Analysis*, Hodder Arnold, London.

Birnbaum A. (2006), Some latent trait models and their use in inferring an examinee's ability, in: Lord F. M., Novick M. R. (eds.), *Statistical Theories of Mental Test Scores*, Addison-Wesley, Reading, MA, 397–424.

Bock R. D., Mislevy R. J. (1982), Adaptive EAP estimation of ability in a microcomputer environment, *Applied Psychological Measurement*, 6, 431–444.

Embretson S. E., Reise S. P. (2000), *Item Response Theory for Psychologists*, Lawrence Erlbaum Associates, Mahwah-New Jersey.

Lazarsfeld P. F., Henry N. W. (1968), *Latent Structure Analysis*, Houghton Mifflin, Boston.

Lord F. M., Novick M. R. (1968), *Statistical Theories of Mental Test Scores*, Addison-Wesley, Reading, MA.

Matteucci M., Mignani S. (2009), *Rapporto tecnico sugli esiti della prova nazionale nell'ambito dell'Esame di Stato al termine del primo ciclo, anno 2007-2008. Analisi delle risposte al test di matematica e italiano: dalle proprietà delle domande alla valutazione degli studenti*, Technical report available at http://www.invalsi.it/EsamiDiStato/documenti/fascicolo_def.pdf

Matteucci M., Mignani S., Ricci R. (2009), Latent variable models to evaluate the final exam in the Italian lower secondary school, in: Ingrassia S., Rocci R. (eds.), *Proceedings of the 7° Meeting of the Classification and Data Analysis Group of the Italian Statistical Society*, CLEUP, Padova, 561–564.

Novick M. R. (1966), The axioms and principal results of classical test theory, *Journal of Mathematical Psychology*, 3, 1–18.

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

van der Linden W. J., Hambleton R. K. (1997), *Handbook of Modern Item Response Theory*, Springer-Verlag, New York.