JOURNAL OF APPLIED QUANTITATIVE METHODS

ISSN 1842-4562
Member of DOAJ

Journal Home > Volume 5, Issue 2, June 30, 2010

The Rasch Model for Evaluating Italian Student Performance

Pdf Version

Ida CAMMINATIELLO
Michele GALLO
Tullio MENINI

Keywords

Programme for International Student Assessment, Italian student performance, Partial Credit Model

Abstract

In 1997 the Organisation for Economic Co-operation and Development (OECD) launched the OECD Programme for International Student Assessment (PISA) for collecting information about 15-year-old students in participating countries. Our study analyse the PISA 2006 cognitive test for evaluating the Italian student performance in mathematics, reading and science comparing the results of different local governments. For this purpose the most proper statistic methodology is Item Response Theory - IRT that collects several models, the simplest is Rasch Model � MR (1960). As the items used in the analysis are both dichotomous that polytomous, we apply Partial Credit Model (PCM).

(top)
(back to issue)

JOURNAL
OF
APPLIED
QUANTITATIVE
METHODS

THE RASCH MODEL FOR EVALUATING ITALIAN STUDENT PERFORMANCE[1]

Ida CAMMINATIELLO
University of Naples Federico II, Naples, Italy

E-mail: camminat@unina.it

Michele GALLO
University of Naples L'Orientale, Naples, Italy

E-mail: mgallo@unior.it

Tullio MENINI
University of Naples L'Orientale, Naples, Italy

E-mail: menini@unior.it

Abstract: *In 1997 the Organisation for Economic Co-operation and Development (OECD) launched the OECD Programme for International Student Assessment (PISA) for collecting information about 15-year-old students in participating countries.*
Our study analyse the PISA 2006 cognitive test for evaluating the Italian student performance in mathematics, reading and science comparing the results of different local governments. For this purpose the most proper statistic methodology is Item Response Theory - IRT that collects several models, the simplest is Rasch Model – MR (1960). As the items used in the analysis are both dichotomous that polytomous, we apply Partial Credit Model (PCM).

Key words: *Programme for International Student Assessment; Italian student performance; Partial Credit Model*

1. Introduction

The prosperity of countries now derives to a large extent from their human capital. This consciousness urges many countries to monitor students' learning. Comparative international assessments can extend and enrich the national picture by providing a larger context within which to interpret national performance. In response to this need, the Organisation for Economic Co-operation and Development (OECD) launched the OECD Programme for International Student Assessment (PISA) in 1997.

The PISA surveys have been designed to collect information about 15-year-old students in participating countries. PISA examines how well students are prepared to meet the challenges of the future, rather than how well they master particular curricula. PISA

J A Q M

Vol. 5
No. 2
Summer
2010

surveys take place every three years. The first survey took place in 2000, the second in 2003 and the third in 2006. For each assessment, one of the three areas (science, reading and mathematics) is chosen as the major domain and given greater emphasis. The remaining two areas, the minor domains, are assessed less thoroughly. In 2000 the major domain was reading; in 2003 it was mathematics and in 2006 it was science. the results of these surveys have been published in a series of reports (OECD, 2001, 2003, 2004, 2007) and a wide range of thematic and technical reports.

In this paper we focus on the PISA 2006 survey. In all countries the survey includes:

- a cognitive test for evaluating the student performance
- a student questionnaire to collect information from students on various aspects of their home, family and school background
- a school questionnaire to collect information from schools about various aspects of organisation and educational provision in schools.

As in previous surveys, additional questionnaire material was developed, which was offered as international options to participating countries. In PISA 2006, two international options were available, the Information Communication Technology (ICT) familiarity and the parent questionnaire.

The PISA 2006 results show wide differences in the performance of countries that participated to the survey. Also the Italian results show performance differences within the country, in particular between local governments and between different schools (INVALSI, 2007).

Our study analyses the cognitive test for evaluating the Italian student performance in reading, mathematics and science, comparing the results of different local governments. Several papers show the measures obtained by students, we are going to focus on measurement instrument for studying:

- the abilities required by PISA 2006 test to which the Italian students are or not able to answer;
- if the students of a local government are scoring better than the students of another local government on an item (Differential Item Functioning, DIF).

For this purpose the most proper statistic methodology is Item Response Theory - IRT (Baker & Kim 2004), that collects several models, the simplest is Rasch Model – MR (1960). As the items used in the analysis are both dichotomous that polytomous, we apply Partial Credit Model (PCM).

2. Rasch model

The aim of the IRT is to test people. Hence, their primary interest is focused on establishing the position of the individual along some not directly observable dimension called latent trait. Because of the many educational applications the *latent trait* is often called *ability*.

The IRT derives the probability of each response as a function of the latent trait and some *item parameters*. The same model is then used to obtain the likelihood of ability as a function of the actually observed responses and, again, the item parameters. The ability value that has the highest likelihood becomes the ability estimate. For this purpose IRT makes the important assumption of *local independence*. This means that the responses given to the separate items in a test are mutually independent *given ability*.

The objective of each IRT model is to predict the probability that a person will give a certain response to a certain item. People can have different levels of ability, and items can have different levels of ability. To keep track of this, we denote the probability of a correct response with $P_{n,s}$: the index *s* refers to the item, and the index *n* refers to the person. When an item allows for more than two options, we denote the probability with $P_{n,s,x}$ where the index *x* refers to the options.

The simplest IRT model is the RM. Rasch's basic idea is that the Models for Measurement make it possible to *measure properly*, and, equally importantly, to validate which data conform to measurement and which does not: Rasch has specified demands for a social sciences measurement to be of the same quality as measurements in the natural sciences and he has then found out exactly which kind of statistical models conform to these specified requirements, namely the Models for Measurement (Rasch, 1968). The conclusion therefore is that a given data set yield measurements in Rasch's well-defined meaning of the word, if, and only if, the data conform to one of the Models for Measurement. So, if the Models for Measurement did not describe the data, then, in certain situations, it is considered better to discard the data than the model.

This view of Rasch's upon data is indeed controversial and quite a contrast to the traditional approach where the statistical model is expanded to fit the data. Closely connected to the Models for Measurement is the concept of specific objectivity, which by and large is the name Rasch chose for his requirements for measurements.

For a dichotomous item the RM has only one item parameter. The probability of a correct response given l'*item parameter* δ_s, and the *person parameter* β_n ,is

$$P_{n,s} = \frac{\exp(\beta_n - \delta_s)}{1 + \exp(\beta_n - \delta_s)} \tag{1}$$

where δ_s characterizes the difficulty of item *s*, and β_n characterizes the ability of examinee *n*.

The literature offers a number of alternative procedures for estimating parameters, including Joint maximum likelihood, Conditional maximum likelihood (CML) and Marginal maximum likelihood (MML). Under appropriate assumptions these solutions are asymptotically equivalent, consistent and multivariate normal (Haberman, 1977; de Leeuw & Verhelst, 1986).

When the items are polytomous with a different number of categories which have not the same distance, the most proper version of the IRT is the Partial Credit Model (PCM) proposed by Wright & Masters (1982). The probability that a subject *n* answers to a item s through the category x ($x = 1,2,\cdots,w,\cdots,M_s$) is calculated by tht formula:

$$P_{nsx} = \frac{\exp\sum_{j=0}^{x}\left[\beta_n - \left(\delta_{sj} + \tau_j\right)\right]}{\sum_{w=0}^{M_s}\exp\sum_{j=0}^{w}\left[\beta_n - \left(\delta_{sj} + \tau_j\right)\right]} \tag{2}$$

δ_{sj} characterizes the difficulty of item *s*, for the threshold *j* and τ_j are category thresholds.

2.1. Rasch diagnostics

In literature there are different tools to evaluate the goodness of fit of the model to observed data. One of the most used is based on the residuals analysis for each individual (or item). The residual can be standardized as follows:

$$z_{ns} = \frac{x_{ns} - E_{ns}}{\sqrt{w_{ns}}}$$

where w_{ns} is the estimated variance of responses reproduced by model, x_{ns} is the response of the individual *n* to the item *s* and E_{ns} is the expected value of the response.

The interpretation of standardised residuals is simple but too analytic because it is referred to each individual or item. For obtaining a synthetic information, the mean value of squared standardised residuals z_{ns}^2 can be calculated: $U_n = \frac{1}{K}\sum_{s=1}^{K} z_{ns}^2$ for each individual where *K* is the number of items and $U_s = \frac{1}{H}\sum_{n=1}^{H} z_{ns}^2$ for each item where *H* the number of individuals.

The expected value of U_n and U_s (outfit or Unweighted Mean Square statistic) is equal to 1. However Linacre proposes different ranges around 1 according to the origin of observed data: for small samples and/or tests with few items, there is a good fit if the statistics is in the range [0.6; 1.4]; otherwise the values should be in [0.8; 1.2]. Anyway values greater than 2 are bad for the measurement .

It can be demonstrated that the outfit statistics is sensitive to big differences between β e δ; for balancing this characteristic it is possible to weigh the squared residuals with the variance, obtaining another synthetic statistics defined INFIT (or Weighted Mean Square statistic): $V_n = \frac{\sum_{s=1}^{K}\left(w_{ns} z_{ns}^2\right)}{\sum_{s=1}^{K}\left(w_{ns}\right)}$ for each individual where *K* is the number of items and $V_s = \frac{\sum_{n=1}^{H}\left(w_{ns} z_{ns}^2\right)}{\sum_{n=1}^{H}\left(w_{ns}\right)}$ for each item where *H* the number of individuals

The infit statistic is sensitive to unexpected behaviour affecting responses to items near the person ability level and the outfit statistic is outlier sensitive, so it is useful to calculate both the statistics.

With reference to the estimations of parameters of RM $\hat{\beta}_n$ and $\hat{\delta}_s$ it is possible to calculate the Standard Error *(SE)*: $SE(\hat{\beta}_n) = \left[\frac{1}{\sum_{s=1}^{K}\left(w_{ns}\right)}\right]^{\frac{1}{2}}$, and $SE(\hat{\delta}_s) = \left[\frac{1}{\sum_{n=1}^{H}\left(w_{ns}\right)}\right]^{\frac{1}{2}}$

Producing a synthesis with respect to the SE of estimations $\hat{\delta}_s$ it is possible to calculate the mean square error: $ME_\delta^2 = \frac{1}{K}\sum_{s=1}^{K}\left[SE\left(\hat{\delta}_s\right)\right]^2$ the squared root of which supplies the mean error of item calibration ME_δ.

The ratio between such value and the squared root of unbiased variance SA_δ gives the separation index: $SI_\delta = \frac{SA_\delta}{ME_\delta}$ where $SA_\delta = \sqrt{\left(S_\delta^2 - ME_\delta^2\right)}$ and $S_\delta^2 = \frac{1}{K}\sum_{s=1}^{K}\delta_s^2$ is the variance of estimations $\hat{\delta}_s$. If the index is far from one, the item are well separated.

In terms of the separation index, realiability index can be expressed as follow: $RE_\delta = \frac{SI_\delta^2}{1+SI_\delta^2} = \frac{SA_\delta^2}{S_\delta^2} = 1 - \frac{ME_\delta^2}{S_\delta^2}$. It has the property that $RE_\delta = 0$ if there is no reproducibility of the measures, $RE_\delta = 1$ if there is perfect reproducibility of the measures, otherwise, $0 \le RE_\delta \le 1$

The goodness of fit can be evaluated graphically by the analysis of Item Characteristic Curves (ICC) and Category Probability Curves (CPC). The ICC of i-*th* item represents the probability of achieving a given score for the item, depending on the parameter value β. The misfit of s-*th* item is observed when one or more points $\hat{p}_{nsx}$ are not on the ICC of the item, where $\hat{p}_{nsx}$ is the probability that individual *n* chooses the category *x* to item *s*, as specified by the Rasch model, with estimated parameters. The CPC provides the probability to choose each of the possible categories according to the difference between ability of the subjects, average difficulty of the item and thresholds among the categories. The thresholds correspond to the measures to which the adjacent categories are equally likely. Compared to the ICC the ordinate represents the expected score for the item, it is obtained by accumulating, for each ability level in abscissa, the product of the estimated probability for each response and the corresponding raw score.

To improve the goodness of fit of a model one can proceed to the elimination of all items (and/or individuals) that do not fit well through an iterative procedure. Often the set of excluded items helps to measure a separate dimension. However, in extreme cases, it can happen it is not possible to identify any set of items consistent with the hypothesis of the Rasch model: this can be caused by a ill calibrated questionnaire or a mixture of individuals apparently belonging to the same population, but in reality related to different populations.

The latter case can be a symptom of a different functioning of the items corresponding to distinct groups of individuals: this phenomenon is called Differential Item Functioning or DIF. More precisely, an item is considered biased if, conditionally to a certain level of ability, the probability of choosing a certain category of response differs systematically between subgroups of individuals (eg., Between males and females). If the presence of DIF is statistically significant, it will be necessary to identify homogeneous groups of individuals that present a good fit.

In literature there are several DIF diagnostics (Glas & Verhelst, 1995), but the most used and implemented in the most commonly used software (Wu, Adams & Wilson, 1998) is based on the residual analysis among the subgroups identified by one or more aggregation variables.

In order to compare the abilities of individuals and the difficulties of the items, one can use the person-item map, a simultaneous graphical representation of both individuals and items. It allows to assess both if an item is more difficult than another one and if an individual is more able than another one.

By convention, the average difficulty of the items in a test is equal to 0 logit: more difficult items than the average difficulty have positive logit values, easier items show negative values. The abilities of individuals are estimated by the model according to the difficulties of the items: a person with an ability equal to 0 logit has a probability equal to 0.5 to successfully pass an item of medium difficulty. More able individuals show positive logit values, less able individuals have negative values. If a person and an item have the same measure on the logit scale, then the person has a probability of 50% to successfully pass the item.

3. Data analysis

3.1. A look at data

The PISA 2006 database includes information on nearly 400,000 students from 57 countries (30 OECD countries and 27 partner countries).

Italy participated to PISA 2006 with a sample of 21,773 students, from 806 schools, stratified by geographical macro-areas (Northwest, Northeast, Central, South, South Island) and fields of study (high schools, technical colleges, vocational schools, secondary schools, vocational training). Moreover, the Italian sample is representative of 11 regions (Basilicata, Campania, Emilia Romagna, Friuli Venezia Giulia, Liguria, Lombardia, Piemonte, Puglia, Sardegna, Sicilia and Veneto) and two autonomous provinces of Bolzano and Trento.

The cognitive test is divided into a variable number of items for each domain. Item formats are multiple-choice, short closed-constructed response, and open-constructed response. Most of the items have only one correct answer (with score 1), then there are some items that allow two correct answers, but with different scores (1 and 2), and some science items that allow three correct answers with scores 1, 2 and 3. In addition, code 9 is used if none of the choices is circled and code 8 if two or more choices are circled. Finally code 7 is reserved for the cases when due to poor printing an item presented to a student is illegible, and therefore the student do not have access to the item.

The mathematics test consist of 48 items (44 have only one correct answer and 4 allow two correct answers). The reading test consist of 28 items (22 admit only one correct answer and the remaining 6 two correct answers. The science test is composed of 192 items.

The descriptive analysis of national and international database shows that each item has about the 69% of 7, so we proceeded to a descriptive analysis for individual and domain. The tables 1, 2 and 3 show the results for student, respectively in mathematics, reading and science at national level.

Table 1. Percentage of illegible items in the mathematics test

Percentage of illegible items for student	Percentage
0 - 50%	45.8%
50 |- 75%	30.7%
75 |- 100%	22.9%

Table 2. Percentage of illegible items in the reading test

Percentage of illegible items for student	Percentage
0 - 50%	53.7%
50 |- 75%	0%
75 |- 100%	46.3%

Table 3. Percentage of illegible items in the science test

Percentage of illegible items for student	Percentage
0 - 50%	15.3%
50 |- 75%	69.4%
75 |- 100%	15.3%

Given the massive presence of missing data, for next analyses we decided to use only the students who have had the opportunity to respond to at least 50% of the items.

3.2 Matematics performance

In this paragraph we analyze Italian student performance in mathematics. The analysis is conducted on the 9963 students who answered at least 50% of the items and 48 items.

The results of the Rasch analysis show an item reliability equal to 1 and a person reliability equal to 0.82, so the test has excellent proprieties of reproducibility. The INFIT and OUTFIT statistics for each item do not present values outside the range [0.6, 1.4], so there is a good fit between data and model for all the items used (Table 4).

Table 4. Item statistics for mathematics

Person: REAL SEP.: 2.12 REL.: .82 ... Item: REAL SEP.: 35.95 REL.: 1.00														
		Item STATISTICS: MISFIT ORDER												
Entry number	Total score	Count	Measure	Model S.E.	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	PTMEA	CORR.	Exact match OBS%	Exact match EXP%	Item	G
20	505	4908	2.37	.05	1.04	1.2	2.05	9.5	A	.28	90.4	90.2	M421Q02T	0
40	4282	4753	-3.37	.05	1.08	2.2	1.89	7.9	B	.30	90.8	91.1	M800Q01	0
22	3733	4860	-2.04	.04	1.12	5.7	1.70	9.9	C	.34	79.4	80.5	M423Q01	0
21	1330	4905	.88	.04	1.13	7.1	1.69	9.9	D	.35	76.3	78.1	M421Q03	0
39	1360	4889	.83	.04	1.10	5.6	1.49	9.9	E	.38	75.7	77.7	M710Q01	0
33	1934	4635	.01	.03	1.17	9.9	1.37	9.9	F	.38	68.0	72.6	M564Q01	0
34	1949	4624	-.01	.03	1.20	9.9	1.35	9.9	G	.36	66.3	72.5	M564Q02	0
12	2630	4892	-.66	.03	1.22	9.9	1.34	9.9	H	.36	63.5	71.7	M305Q01	0
36	2900	4877	-.99	.03	1.14	8.8	1.23	7.5	I	.43	68.5	73.5	M598Q01	0
1	3599	4998	-1.71	.04	1.07	3.8	1.21	5.3	J	.42	76.0	77.8	M033Q01	0
8	2198	4874	-.18	.03	1.11	7.6	1.19	6.5	K	.43	69.1	72.4	M273Q01T	0
10	3640	4923	-1.90	.04	1.00	-.2	1.19	4.0	L	.47	80.9	79.6	M302Q02	0
48	1297	4664	.88	.04	1.12	6.2	1.14	3.1	M	.39	74.3	77.6	M833Q01T	0
18	2124	4850	-.10	.03	1.09	5.9	1.13	4.6	N	.45	69.4	72.6	M420Q01T	0
5	1048	4955	1.86	.03	.98	-.6	1.10	1.2	O	.48	84.8	82.9	M155Q03T	0
17	1949	4910	.13	.03	1.02	1.2	1.10	3.4	P	.47	72.5	72.7	M411Q02	0
29	3670	4997	-1.80	.04	1.01	.3	1.09	2.3	Q	.46	79.0	78.6	M474Q01	0
15	2088	4865	-.05	.03	1.05	3.7	1.08	2.7	R	.47	70.8	72.7	M408Q01T	0
6	2460	4944	-.42	.03	1.04	2.7	1.05	2.0	S	.47	70.1	71.7	M155Q04T	0
37	2384	4744	-.45	.03	1.02	1.8	1.03	1.4	T	.48	71.0	71.6	M603Q01T	0
47	1265	4766	.94	.04	1.03	1.8	.92	-1.9	U	.45	76.6	78.5	M828Q03	0
4	4299	4988	-.13	.02	1.03	1.3	.98	-.4	V	.64	57.8	58.3	M155Q02T	0
32	3162	4806	-1.33	.04	.99	-.3	1.03	.7	W	.50	75.4	75.4	M559Q01	0
27	1015	4775	1.84	.03	1.01	.4	.98	-.2	X	.46	83.7	82.9	M462Q01T	0
35	1934	4676	.03	.03	.93	-4.8	1.00	.0	x	.53	76.0	72.7	M571Q01	0
31	2919	4875	-.99	.03	.99	-.4	1.00	.0	w	.49	73.3	72.6	M496Q02	0
42	1820	4806	.24	.03	.99	-1.0	.93	-2.4	v	.50	74.1	73.7	M810Q01T	0
7	1429	4853	.73	.04	.97	-1.5	.98	-.4	u	.48	77.5	77.0	M192Q01T	0
26	3151	4885	-1.26	.03	.97	-2.1	.96	-1.2	t	.52	75.8	74.9	M447Q01	0
2	1575	4714	.52	.04	.96	-2.6	.88	-4.0	s	.50	75.5	74.7	M034Q01T	0
16	2338	4943	-.29	.03	.96	-3.2	.94	-2.7	r	.53	73.1	71.7	M411Q01	0
46	2255	4775	-.28	.03	.95	-3.5	.92	-3.1	q	.54	74.1	72.3	M828Q02	0
38	1537	4732	.55	.04	.95	-3.2	.87	-4.2	p	.52	76.3	75.6	M603Q02T	0
25	186	4814	3.61	.08	.93	-1.1	.74	-2.1	o	.29	96.2	96.2	M446Q02	0
45	940	4802	1.45	.04	.93	-3.2	.75	-4.6	n	.48	83.2	82.8	M828Q01	0
30	2029	4880	.01	.03	.92	-5.8	.87	-5.4	m	.55	75.0	72.7	M496Q01T	0
3	3043	4957	-1.06	.03	.91	-6.1	.85	-5.6	l	.55	76.4	73.3	M155Q01	0
9	4607	4937	-4.03	.07	.91	-1.9	.65	-3.5	k	.43	94.8	94.3	M302Q01T	0
24	2937	4838	-1.03	.03	.90	-7.1	.86	-4.9	j	.56	77.8	73.6	M446Q01	0
11	1086	4923	1.24	.04	.89	-5.3	.74	-5.3	i	.50	82.6	81.0	M302Q03	0
43	2724	4741	-.84	.03	.89	-7.4	.84	-6.0	h	.57	76.4	72.9	M810Q02T	0
41	1029	4891	1.31	.04	.89	-5.4	.70	-6.7	g	.50	82.5	81.4	M803Q01T	0
28	1203	4763	1.03	.04	.89	-5.9	.75	-5.8	f	.53	81.7	79.2	M464Q01T	0
23	1849	4825	.23	.03	.87	-9.4	.79	-8.2	e	.57	78.0	73.0	M442Q02	0
44	1281	4733	1.65	.03	.84	-5.5	.54	-7.3	d	.56	82.5	78.8	M810Q03T	0
14	624	4768	2.04	.05	.84	-5.7	.55	-7.5	c	.49	88.9	87.8	M406Q02	0
13	1138	4814	1.12	.04	.83	-8.6	.68	-8.0	b	.55	83.6	80.2	M406Q01	0
19	2542	4914	-.56	.03	.82	-9.9	.74	-9.9	a	.62	79.3	72.4	M421Q01	0
MEAN	2145.8	4838.8	.00	.04	.99	-.3	1.04	.2			77.2	77.0		
S.D.	1042.8	93.6	1.41	.01	.10	5.3	.32	5.8			7.5	6.6		

The measures, the abilities of the students and the difficulties of the items can be displayed graphically through the person-item map (Figure 1). It is noted that in the central part of the graph there is most of the students represented on the left by # (each # represents 48 students), and most of the items represented on the right by label of the item. The test is quite broad, though slightly upon the mean level of students, the items and the students are quite well approximated by a normal distribution The item M446Q02 is the most difficult, on the contrary the items M800Q01 and M302Q01T are the simplest. At the bottom of the chart there is a small group of students, at which there are no items able to measure their ability.

```
             Persons - MAP - Items
                     <more>|    <rare>
 5                        .  +
                             |
                             |
                             |
                             |
                          .  |
 4                        .  +
                          .  |
                          .  |    M446Q02
                          .  |
                          .  |
                          .  |
 3                        .  +
                             | T
                          .  |
                         .#  |
                         .# T|    M421Q02T
                         .#  |
 2                      .##  +    M406Q02
                         .#  |    M155Q03T  M462Q01T
                       .###  |    M810Q03T
                       .###  |    M828Q01
                     .#####  | S  M803Q01T
                    .######  |    M302Q03   M406Q01
 1                   .#### S +    M464Q01T  M828Q03
                      .####  |    M421Q03   M710Q01   M833Q01T
                  .########  |    M192Q01T
                   .#######  |    M034Q01T  M603Q02T
                .##########  |
                    .######  |    M411Q02   M442Q02   M810Q01T
 0            .############  + M  M408Q01T  M496Q01T  M564Q01   M564Q02   M571Q01
                  .########  |    M155Q02T  M273Q01T  M420Q01T
                  .########  |    M411Q01   M828Q02
                .########## M|    M155Q04T  M421Q01   M603Q01T
              .############  |    M305Q01
                   .#######  |    M810Q02T
-1              .##########  +    M155Q01   M446Q01   M496Q02   M598Q01
                  .########  |
                   .#######  | S  M447Q01   M559Q01
                  .########  |
                    .######  |    M033Q01
                     .##### S|    M302Q02   M474Q01
-2                      .##  +    M423Q01
                    .######  |
                       .###  |
                       .###  |
                         .#  |
                        .##  | T
-3                       .#  +
                        .## T|
                          .  |    M800Q01
                          #  |
                          .  |
                          .  |
-4                       .#  +    M302Q01T
                          .  |
                             |
                          .  |
                             |
                             |
-5                       .#  +
                     <less>|    <frequ>
 EACH '#' IS 48.
```

Figure 1. Person-item map for mathematics

In order to verify that the thresholds are ordered and between them there is a suitable distance, we show the CPCs. To not bore the reader, the figure 1 shows only the CPCs of the four items that allow two correct answers with score 1 and 2. It is easy to check that, for each them, the curve of probability of category 0 meets, first, the curve of probability of category 2 and, then, that of category 1. The category 1, therefore, is never the most likely. To improve the interpretation of the measures it could be appropriate a pooling of the categories 0 and 1 of these items.

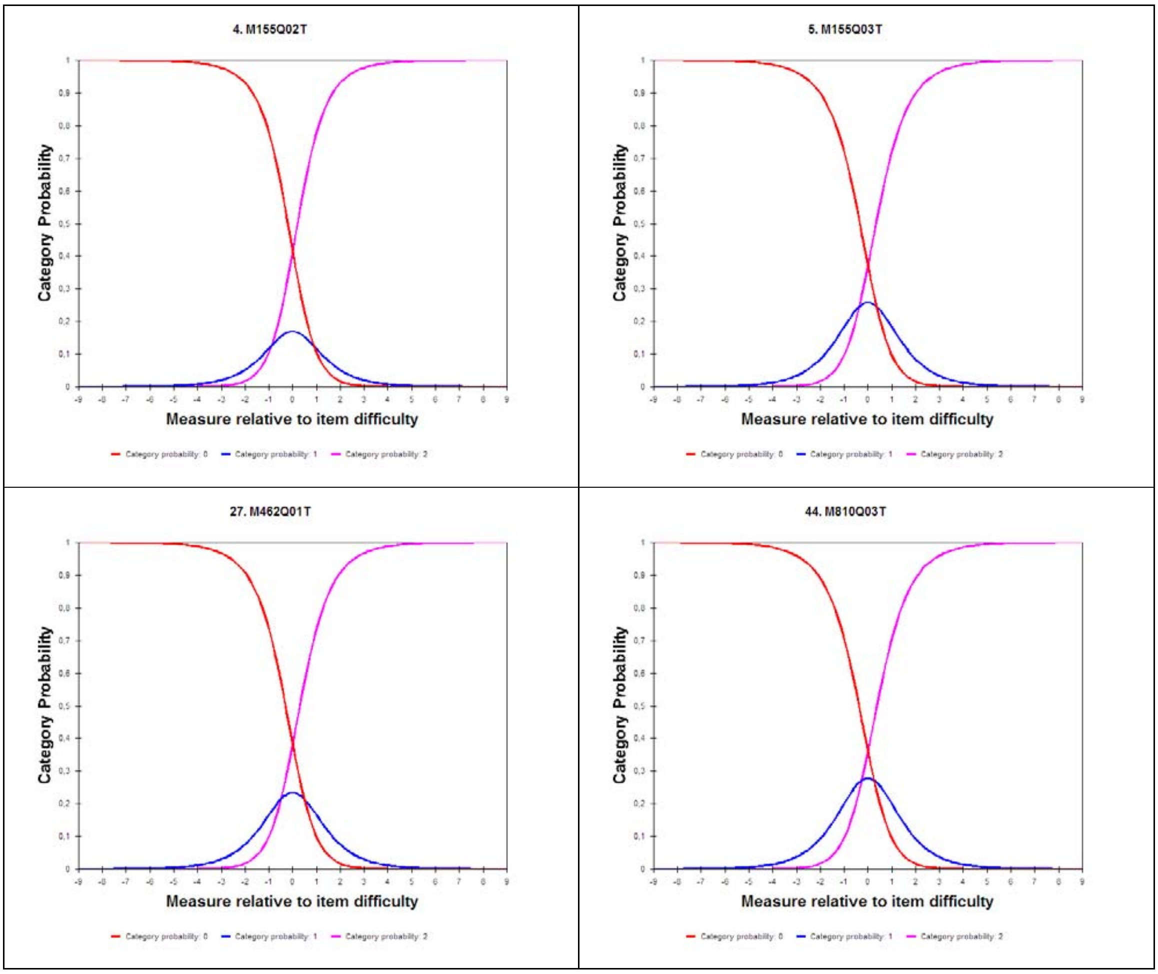

Figure 2. Mathematics performance: CPCs of the four items that allow two correct answers

The final step of the analysis is the comparison of estimates among two or more groups to examine whether the items have a significantly different functioning. This phenomenon is called Differential Item Functioning (DIF). In our case it is interesting to examine the functioning of items among the various Italian regions.

For this purpose, the table 5 allows us to test the hypothesis that the items have the same functioning among the several Italian regions. The table shows that 29 items have a statistically different functioning among the various Italian regions at a significance level of 5%. These items are in red in Table 5.

Table 5. Mathematics performance: DIF among Italian regions

Person CLASSES	SUMMARY DIF CHI-SQUARE	D.F.	PROB.	Item Number	Item Name
14	28.1544	13	.0086	1	M033Q01
14	22.7786	13	.0444	2	M034Q01T
14	13.4938	13	.4104	3	M155Q01
14	33.1855	13	.0016	4	M155Q02T
14	17.7838	13	.1659	5	M155Q03T
14	25.8611	13	.0177	6	M155Q04T
14	29.5869	13	.0054	7	M192Q01T
14	30.0954	13	.0046	8	M273Q01T
14	18.9098	13	.1259	9	M302Q01T
14	5.8639	13	.9510	10	M302Q02
14	24.3117	13	.0284	11	M302Q03
14	26.5099	13	.0145	12	M305Q01
14	20.1734	13	.0909	13	M406Q01
14	28.4531	13	.0078	14	M406Q02
14	32.6207	13	.0019	15	M408Q01T
14	7.9870	13	.8444	16	M411Q01
14	27.3574	13	.0111	17	M411Q02
14	31.8679	13	.0025	18	M420Q01T
14	60.1173	13	.0000	19	M421Q01
14	43.5071	13	.0000	20	M421Q02T
14	37.5156	13	.0003	21	M421Q03
14	69.3453	13	.0000	22	M423Q01
14	38.4619	13	.0002	23	M442Q02
14	36.1250	13	.0006	24	M446Q01
14	14.8873	13	.3144	25	M446Q02
14	17.9794	13	.1583	26	M447Q01
14	58.3753	13	.0000	27	M462Q01T
14	14.3670	13	.3485	28	M464Q01T
14	8.9987	13	.7730	29	M474Q01
14	15.2868	13	.2898	30	M496Q01T
14	13.5657	13	.4051	31	M496Q02
14	29.7048	13	.0052	32	M559Q01
14	33.9595	13	.0012	33	M564Q01
14	10.7980	13	.6277	34	M564Q02
14	13.0336	13	.4452	35	M571Q01
14	36.8692	13	.0004	36	M598Q01
14	16.8599	13	.2058	37	M603Q01T
14	21.0081	13	.0727	38	M603Q02T
14	18.6118	13	.1356	39	M710Q01
14	63.1808	13	.0000	40	M800Q01
14	30.2286	13	.0044	41	M803Q01T
14	24.2321	13	.0290	42	M810Q01T
14	35.2862	13	.0008	43	M810Q02T
14	19.3109	13	.1137	44	M810Q03T
14	81.3797	13	.0000	45	M828Q01
14	13.0175	13	.4464	46	M828Q02
14	23.3072	13	.0381	47	M828Q03
14	24.1007	13	.0302	48	M833Q01T

To understand the magnitude of the differences between the regions is interesting to look at figure 3. This graph shows the difficulty of each item for each region. From the figure 3. it would seem that there are no appreciable differences of the items between different regions. However, a small value of DIF could be statistically significant, while a large value of DIF could be not statistically significant, so it is important to look at the chi-square test above illustrated (table 5.).

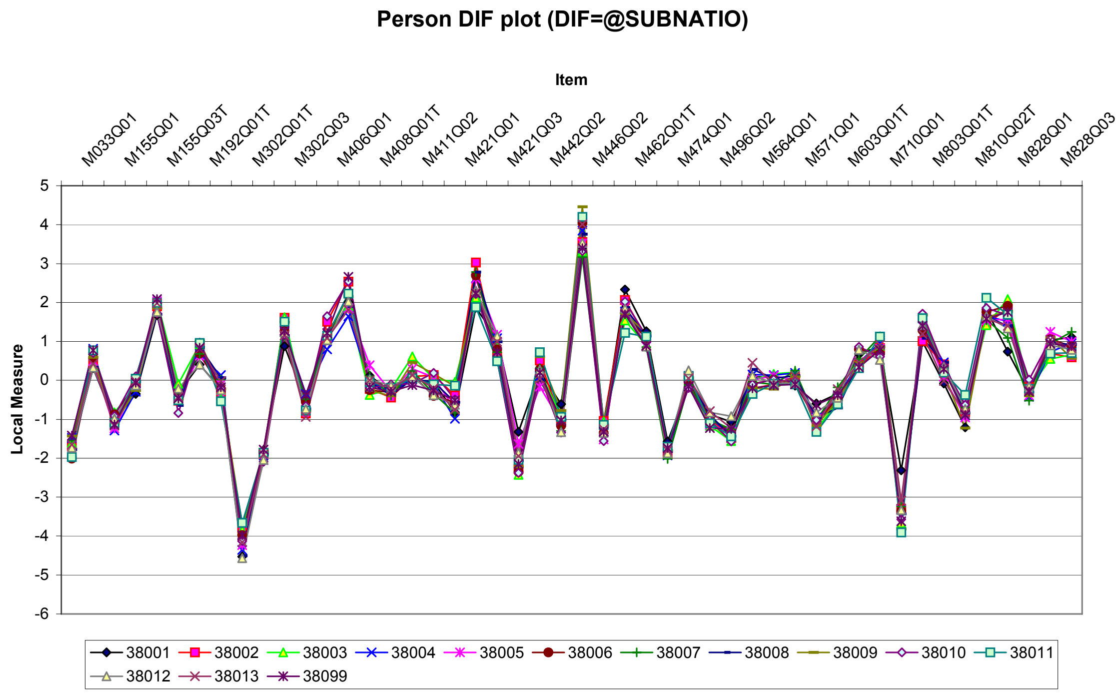

Figure 3. Mathematics performance: difficulty of each item for each region.

3.3 Reading performance

In this paragraph we analyze Italian student performance in reading. The analysis is conducted on the 11686 students who answered at least 50% of the items and 28 items.

The results of the Rasch analysis show an item reliability equal to 1 and a person reliability equal to 0.78, so the test has good proprieties of reproducibility. From the table 6. we can observe that the INFIT and OUTFIT statistics present values outside the range [0.6, 1.4] for the following four items: R111Q06B, R067Q04, R227Q02T R067Q05. This could be due to a different functioning of the items among the various Italian regions. This hypothesis will be verified by analysis of DIF. Moreover, it would be appropriate to remove or replace these items because they could distort the measures obtained. However, we prefer not to make these changes to remain faithful to the test calibrated at international level. The stakeholders can focus on the contents of such items to address the educational proposals towards the disciplinary facets which are more problematic.

For this purpose the PISA compendium has been published. It gathers the PISA tests that have been issued in various editions and administered in Main Studies, i.e. those which have been published and will not be reused in subsequent cycles. The compendium is divided into three parts: Reading, Math and Science. Each test is accompanied by the description of items, by the guide for the correction of responses and by the data on student responses at different levels: average of the OECD countries, the national percentages, the percentages for macro-areas. The original numbering of the items has been left, so it is easy to establish the correspondence between item and content.

Table 6. Item statistics for Reading

Person: REAL SEP.: 1.88 REL.: .78 ... Item: REAL SEP.: 40.43 REL.: 1.00

Item STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	PTMEA CORR.	Exact match OBS%	Exact match EXP%	Item
8	1598	6551	2.16	.03	.86	-8.7	.80	-4.6	.44	80.3	76.9	R102Q04A
20	2285	6497	1.54	.03	.81	-9.9	.78	-6.7	.52	78.0	71.8	R220Q01
9	2318	6508	1.51	.03	.84	-9.9	.81	-5.8	.50	76.7	71.7	R102Q05
12	2229	6360	1.38	.03	1.19	9.9	1.66	9.9	.30	65.8	72.9	R104Q02
13	2284	6324	1.32	.03	1.32	9.9	1.17	4.8	.48	71.5	72.6	R104Q05
17	2594	6625	1.31	.03	.84	-9.9	.83	-5.6	.51	75.9	71.1	R219Q01E
25	2792	6631	.95	.03	1.21	9.9	1.67	9.9	.32	63.0	71.8	R227Q01
2	2889	6524	.86	.03	.89	-7.2	.91	-3.3	.52	72.5	71.6	R055Q02
15	3370	6561	.49	.03	1.16	9.3	1.08	3.4	.56	70.6	71.6	R111Q02B
18	3880	6643	.36	.03	.74	-9.9	.72	-9.9	.58	76.7	71.4	R219Q01T
27	3614	6619	.32	.03	.73	-9.9	.69	-9.9	.62	78.2	71.7	R227Q03
21	3855	6398	.32	.03	.76	-9.9	.86	-6.4	.55	76.1	71.5	R220Q02B
14	3776	6590	.20	.03	.75	-9.9	.78	-9.9	.60	78.0	72.0	R111Q01
3	3771	6456	.18	.03	.75	-9.9	.72	-9.9	.59	77.2	71.9	R055Q03
22	4181	6387	.07	.03	.76	-9.9	.82	-8.7	.53	76.8	72.2	R220Q04
24	4363	6369	-.07	.03	.74	-9.9	.79	-9.9	.53	77.2	72.5	R220Q06
19	4513	6618	-.10	.03	.69	-9.9	.71	-9.9	.57	78.9	72.6	R219Q02
28	4233	6590	-.14	.03	.69	-9.9	.69	-9.9	.61	79.2	72.5	R227Q06
4	4276	6448	-.20	.03	.59	-9.9	.57	-9.9	.66	83.1	72.5	R055Q05
23	4971	6377	-.52	.03	.53	-9.9	.55	-9.9	.60	84.3	73.4	R220Q05
11	4722	6412	-.55	.03	.70	-9.9	.72	-9.9	.53	79.1	73.0	R104Q01
16	5262	6541	-.91	.03	2.45	9.9	2.47	9.9	.65	34.7	72.9	R111Q06B
10	5591	6501	-.92	.03	.51	-9.9	.53	-9.9	.51	86.5	73.5	R102Q07
1	5308	6537	-.94	.03	.59	-9.9	.62	-9.9	.53	83.1	72.9	R055Q01
5	5948	6609	-1.12	.03	.47	-9.9	.50	-9.9	.47	87.9	73.2	R067Q01
6	7343	6594	-2.08	.03	2.01	9.9	2.06	9.9	.56	43.8	68.6	R067Q04
26	7558	6627	-2.45	.03	1.57	9.9	1.61	9.9	.58	54.2	67.7	R227Q02T
7	8707	6570	-2.98	.03	2.29	9.9	2.32	9.9	.65	41.5	66.8	R067Q05

In order to compare the student ability and the item difficulty, we present the person-item map (Figure 5). The test is significantly upon the mean level of students, in fact, there is a large group of students for who there are no items calibrated on their ability level, on the contrary, there are some very difficult items (R102Q04A, R102Q05, R220Q01, R104Q02, R104Q05 , R219Q01E, R227Q01, R055Q02) at which they are no students or there is a very small number. The students are asymmetrically distributed.

The analysis of the CPCs does not show problematic aspects: the thresholds are ordered and their distance is sufficient (Figure 9). Indeed Linacre (1999) indicates that the thresholds should grow at least 1.4 logit for different categories, but not more than 5 logit to ensure continuity of the variable.

Finally, we examine the functioning of items among the various Italian regions. The table 7 shows that almost all the items (red-ink in the table 7) have a statistically different functioning among the various Italian regions at a significance level of 5%, so it would be desirable to identify a battery of items that can operate in not statistically different way between the Italian regions. The DIF analysis seems to verify the hypothesis that a different functioning of the items among the various Italian regions lead to their poor fit. In fact, the four items that have a bad fit also have a significant DIF.

The magnitude of the differences between regions is represented in Figure 6, where one can see that the variability of item difficulty is greater when the item has a different functioning between the regions. For example, Bolzano (38001) seems the more different region than the others.

```
Persons - MAP - Items
                    <more>     |     <rare>
  3                            +
                               |
                               |
                               |  T
                               |     R102Q04A
  2                            +
                               |
                               |     R102Q05   R220Q01
                         .     |     R104Q02   R104Q05   R219Q01E
                            T  |  S
  1                      .     +     R227Q01
                         .     |     R055Q02
                       .##     |
                       .##     |     R111Q02B  R219Q01T  R220Q02B  R227Q03
                     .####     |     R055Q03   R111Q01
  0             .#########     +  M  R219Q02   R220Q04   R220Q06
                         #  S  |     R055Q05   R227Q06
             .############     |
             .############     |     R104Q01   R220Q05
                        .#     |
 -1           .###########     +     R055Q01   R102Q07   R111Q06B
             .############     |  S  R067Q01
                        .#     |
               .##########  M  |
                    .#####     |
 -2                 .#####     +     R067Q04
                 .########     |
                     .####     |  T  R227Q02T
                      .###     |
                     .####     |
 -3                   .###  S  +     R067Q05
                      .###     |
                       .##     |
                      .###     |
                         .     |
 -4                    .##     +
                      .###     |
                         .  T  |
                        .#     |
                         .     |
 -5                      .     +
                       .##     |
                         .     |
                        .#     |
                               |
 -6                            +
                         .     |
                               |
                               |
                               |
 -7                    .##     +
                    <less>     |     <frequ>
EACH   '#' IS 81.
```

Figure 5. Person-item map for Reading

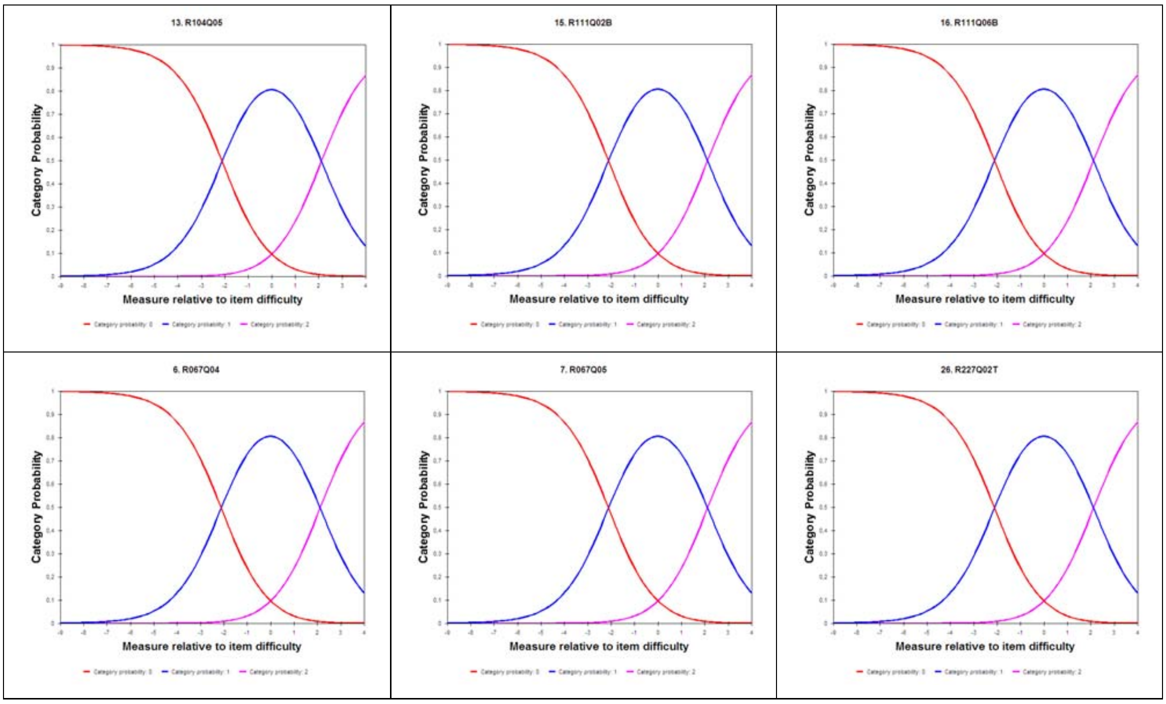

Figure 4. Reading performance: CPCs of the six items that allow two correct answers

Table 7. Reading performance: DIF among Italian regions

Person CLASSES	SUMMARY DIF CHI-SQUARE	D.F.	PROB.	Item	
				Number	Name
14	33.2730	13	.0015	1	R055Q01
14	37.5273	13	.0003	2	R055Q02
14	28.6081	13	.0074	3	R055Q03
14	18.8289	13	.1285	4	R055Q05
14	127.210	13	.0000	5	R067Q01
14	81.1536	13	.0000	6	R067Q04
14	49.2266	13	.0000	7	R067Q05
14	80.4530	13	.0000	8	R102Q04A
14	28.2015	13	.0085	9	R102Q05
14	134.343	13	.0000	10	R102Q07
14	27.0227	13	.0123	11	R104Q01
14	37.8045	13	.0003	12	R104Q02
14	25.2743	13	.0212	13	R104Q05
14	14.0466	13	.3705	14	R111Q01
14	33.8958	13	.0012	15	R111Q02B
14	196.096	13	.0000	16	R111Q06B
14	76.4062	13	.0000	17	R219Q01E
14	63.5525	13	.0000	18	R219Q01T
14	31.4878	13	.0029	19	R219Q02
14	90.7701	13	.0000	20	R220Q01
14	11.4726	13	.5713	21	R220Q02B
14	26.6078	13	.0141	22	R220Q04
14	18.8257	13	.1286	23	R220Q05
14	33.4510	13	.0015	24	R220Q06
14	32.9223	13	.0017	25	R227Q01
14	34.3976	13	.0010	26	R227Q02T
14	12.7273	13	.4691	27	R227Q03
14	13.1172	13	.4388	28	R227Q06

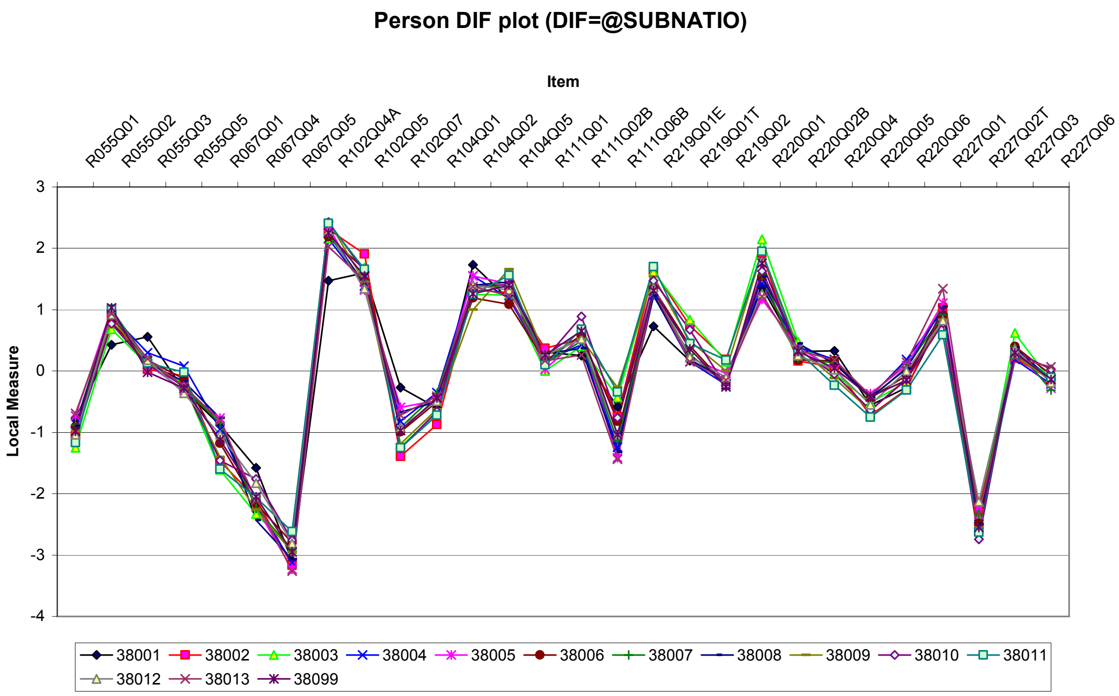

Figure 6. Reading performance: difficulty of each item for each region

3.4. Science performance

The analysis Italian student performance in science is conducted on the 3311 (15,3%) students who answered at least 50% of the items . Though the results of the Rasch analysis show an item reliability equal to 1 and a person reliability equal to 0.92, 22 items present the INFIT and OUTFIT statistics outside the range [0.6, 1.4]. These items are in table 8.

Table 8. The INFIT and OUTFIT statistics of 22 science items outside the range [0.6, 1.4]

ENTRY NUMBER	TOTAL SCORE	COUNT	Measure	Model S.E.	Infit		Outfit		PTMEA CORR.	Exact match		Item
					MNSQ	ZSTD	MNSQ	ZSTD		OBS%	EXP%	
94	745	1551	1.83	.05	1.92	9.9	1.83	9.9	.35	54.9	60.0	S519Q01
56	868	1642	1.66	.05	2.22	9.9	2.09	9.9	.42	41.4	58.8	S447Q05
2	1085	1669	1.26	.04	1.80	9.9	1.78	9.9	.38	37.8	58.6	S114Q04T
76	1297	1669	.87	.04	1.50	9.9	1.51	9.9	.32	43.8	61.3	S485Q05
86	1291	1466	.54	.04	2.29	9.9	2.27	9.9	.45	10.0	63.0	S498Q04
59	1637	1642	.23	.04	1.79	9.9	1.80	9.9	.38	27.8	63.4	S465Q01
149	1908	1653	-.19	.04	1.75	9.9	1.76	9.9	.25	45.4	61.8	S521QNB
148	2014	1651	-.36	.04	1.63	9.9	1.63	9.9	.32	47.4	60.1	S521QNA
119	1975	1606	-.38	.04	1.72	9.9	1.73	9.9	.31	45.1	59.5	S438QNB
109	2006	1605	-.43	.04	1.85	9.9	1.85	9.9	.37	44.2	58.9	S413QNC
118	2119	1609	-.59	.04	1.73	9.9	1.76	9.9	.27	42.0	57.0	S438QNA
120	2333	1607	-.91	.04	1.54	9.9	1.56	9.9	.36	43.6	53.4	S438QNC
116	2473	1626	-1.07	.04	1.46	9.9	1.48	9.9	.34	41.0	51.7	S437QNB
135	2547	1649	-1.12	.04	1.50	9.9	1.50	9.9	.35	40.0	51.4	S485QNC
104	2615	1642	-1.23	.04	1.65	9.9	1.66	9.9	.34	37.7	50.9	S408QNA
136	2356	1444	-1.33	.04	1.51	9.9	1.52	9.9	.29	41.4	50.9	S498QNA

ENTRY NUMBER	TOTAL SCORE	COUNT	Measure	Model S.E.	Infit		Outfit		PTMEA CORR.	Exact match		Item
					MNSQ	ZSTD	MNSQ	ZSTD		OBS%	EXP%	
114	2657	1618	-1.33	.04	1.41	9.9	1.43	9.9	.44	43.3	50.8	S428QNC
129	2738	1654	-1.37	.04	1.54	9.9	1.57	9.9	.28	39.2	50.5	S476QNC
153	2760	1657	-1.39	.04	1.44	9.9	1.47	9.9	.30	40.3	50.4	S527QNA
150	2343	1396	-1.43	.04	1.49	9.9	1.51	9.9	.35	42.1	50.6	S524QNA
115	2830	1626	-1.54	.04	1.46	9.9	1.47	9.9	.28	42.7	50.2	S437QNA
117	3010	1628	-1.77	.04	1.41	9.9	1.42	9.9	.40	43.6	50.2	S437QNC

The person-item map (Figure 7.) shows that items are distributed into two main blocks: half (maybe more than half) of the items are more difficult than the average, about half of the items are easier than the average, while items of average difficulty (by convention, the average difficulty of items in a test is set equal to 0) are missing. Looking at the distribution of the items one feels that the test measure two different dimensions. Looking at the distribution of individuals, it is easy to see that it is normal and symmetric with respect to -1 rather than 0. Consequently, it would be appropriate to introduce items of average difficulty and to reduce the number of easy and difficult items.

```
Persons - MAP - Items
              <more> |  <rare>
 4                   + T
                     |  S527Q01T
                     |  S114Q05T
                     |  S458Q01
 3                   +  S425Q04   S524Q07
                     |  S326Q04T  S425Q03   S519Q03
                     |  S131Q04T  S268Q02T  S269Q04T  S304Q03A  S408Q03
                        S413Q04T  S447Q02   S498Q03   S510Q04T
                     |  S213Q01T  S269Q03T  S408Q04T  S413Q06   S425Q02
                        S438Q03T  S447Q04   S465Q04   S495Q01T
 2                   + S S304Q01  S408Q05   S416Q01   S421Q01   S428Q05
                        S437Q03   S493Q01T  S495Q03   S498Q02T  S510Q01T
                        S514Q03
                     |  S114Q03T  S131Q02T  S269Q01   S304Q03B  S447Q05
                        S477Q04   S478Q01   S493Q05T  S495Q02T  S514Q04
                        S519Q01   S519Q02T  S524Q06T
                     |  S268Q06   S326Q01   S326Q02   S326Q03   S415Q08T
                        S421Q03   S426Q03   S426Q07T  S437Q04   S438Q02
                        S447Q03   S465Q02   S466Q05   S478Q02T  S485Q02
                        S495Q04T  S521Q02   S527Q03T
                   . |  S114Q04T  S268Q01   S304Q02   S408Q01   S425Q05
                        S428Q03   S437Q01   S437Q06   S458Q02T  S476Q03
                        S478Q03T  S485Q03   S508Q02T  S508Q03   S527Q04T
 1                 . +  S213Q02   S413Q05   S415Q02   S415Q07T  S428Q01
                        S466Q01T  S466Q07T  S476Q01   S476Q02   S477Q02
                        S514Q02
                   . |  S256Q01   S426Q05   S438Q01T  S477Q03   S485Q05
                        S493Q03T
                  .T |  S498Q04   S521Q06
                  .# |  S465Q01
```

```
  0           .####    + M
          .#######S    |   S521QNA  S521QNB
       .###########    |   S413QNC  S438QNA  S438QNB
.      ###########M    |
 -1    .###########    +   S408QNB  S408QNC  S437QNB  S438QNC  S485QNC
          .########    |   S408QNA  S413QNA  S413QNB  S428QNB  S428QNC
                           S476QNC  S498QNA  S508QNB  S514QNB
            .#####S    |   S408QSC  S428QNA  S437QNA  S456QNA  S466QNB
                           S508QNA  S508QNC  S514QNA  S524QNA  S524QNB
                           S524QNC  S527QNA
                ###    |   S408QSB  S416QNA  S426QSB  S437QNC  S438QSB
                           S456QNB  S485QNA  S485QNB  S498QNB  S498QNC
                           S527QNC
 -2             .#T    + S S408QSA  S425QSC  S456QNC  S456QSB  S456QSC
                           S466QNA  S476QNA  S476QNB  S478QNA  S478QNB
                           S478QNC  S514QNC
                  .    |   S416QSA  S416QSB  S438QSC  S466QNC  S477QSB
                           S485QSB  S485QSC  S519QNB  S519QNC  S527QNB
                  .    |   S416QNB  S421QSC  S426QSA  S438QSA  S465QSA
                           S465QSB  S477QSC  S498QSA  S498QSB
                  .    |   S425QSB  S426QSC  S456QSA  S519QNA  S519QSB
                           S527QSB
 -3               .    +   S416QSC  S421QSA  S476QSA  S527QSC
                  .    |   S425QSA  S477QSA  S519QSA  S519QSC
                  .    |
                  .    |   S476QSB
 -4               .    + T S476QSC
                  .    |
                       |
                  .    |
 -5               .    +
                       |
                  .    |
                       |
 -6                    +
             <less>    |   <frequ>
 EACH '#' IS 47.
```

Figure 7. Person-item map for science

As we wrote above about reading performance, it would be appropriate to remove or replace these items which present the INFIT and OUTFIT statistics outside the range [0.6, 1.4], because they could distort the measures obtained. However, we prefer not to make these changes to remain faithful to the test calibrated at international level. The stakeholders can focus on the contents of such items to address the educational proposals towards the disciplinary facets which are more problematic.

In the light of these results we believe it is not necessary to show the CPC curves and to compare the different Italian regions.

4. Final remarks

In this paper we analyzed the cognitive test for evaluating the Italian student performance in reading, mathematics and science, comparing the results of different local governments.

The descriptive analysis of national and international database showed that each item has about the 69% of 7, so we proceeded to a descriptive analysis for individual and domain in order to identify students who have had the opportunity to respond to at least 50% of the items.

Given the considerable presence of missing data, we could opt for different strategies, such as the use of algorithms for estimating the missing data or the analysis of the only available data. Unlike the strategy applied by the OECD that decided to estimate the missing data, we chose to include in the analysis only the students who have had the opportunity to respond to at least 50% of the items.

The results for mathematics performance can be summarized as follows:

- the test has excellent proprieties of reproducibility;
- All the items show a good fit (the INFIT and OUTFIT statistics for each item do not present values outside the range);
- The item map shows that the test is quite broad, though slightly upon the mean level of students;
- There are some items that have a significantly different functioning between the Italian regions at the 5% level.

For reading performance, 4 items have a bad fit, the test is significantly upon the mean level of students and there are some items that have a significantly different functioning between the Italian regions at the 5% level.

For science, the situation is very controversial. The analysis of Italian student performance was conducted on a small sample because of the considerable presence of missing data. We found many items have a bad fit. The person-item map shows it would be appropriate to introduce items of average difficulty.

In the light of the results obtained in the three domain, the stakeholders should address the educational proposals towards the disciplinary facets of mathematics, reading and science which are more problematic in order to reduce the differences between the regions and to improve the Italian student performance. It would also be interesting to compare the results obtained in different domain by using algorithms for estimating the missing data.

References

1. S. Bacci **I modelli di Rasch nella valutazione della didattica universitaria**, Statistica Applicata, vol. 18, n. 1, 2006, pp. 1-40.
2. F. Baker and S. Kim **Item response theory. Parameter estimation techniques**, Dekker, New York, 2004
3. Trevor G. Bond and Christine M. Fox **Applying the Rasch Model**, Lawrence Erlbaum, Mahwah, NJ
4. A. Glas and N. Verhelst **Tests of fit for polytomous rasch models**, in G. H. Fischer e I. W. Molenaar, eds, Rasch models. Foundations, recent developments, and applications. Springer-Verlag, 1995, pp. 325–352
5. INVALSI **Compendio Prove PISA**, www.invalsi.it, 2009
6. INVALSI **Le competenze in scienze lettura e matematica degli studenti quindicenni**, Armando Editore, 2007
7. J. M. Linacre **Many-Facet Rasch Measurement**, Chicago, MESA Press, 1989

8. J. M. Linacre **Facets Rasch measurement computer program**, Chicago, Winsteps.com, 2004
9. G. N. Masters **A Rasch model for partial credit scoring**, «Psychometrika», vol. 47, 1982
10. R. Miceli **Percorsi di ricerca e analisi dei dati**, Torino, Bollati Boringhieri, 2001
11. S. Mignani and R. Ricci **Il ruolo del metodo statistico nel progetto PISA**, «Induzioni», vol. 30, 2005, pp. 59-73
12. OECD, **PISA 2006: Science Competencies for Tomorrow's World**, Paris, 2007, pp. 283-326
13. OECD, **PISA 2006 Technical Report**, Paris, 2009, pp. 143-156
14. B. D. Wright and G. N. Masters **Rating Scale Analysis**, Chicago, MESA Press, 1982
15. B.D. Wright and J.M. Linacre **Observations are Always Ordinal: Measures, However, Must be Interval**, «Archives of Physical Medicine and Rehabilitation», Vol. 70, 1989
16. M. Wu, R. Adams and M. Wilson **Acer Conquest. Generalised item response modelling software**, Acer Press, 1998

[1] **Acknowledgements:** The present paper has been carried out within the Nuval project: "Analisi e valutazione delle politiche di sviluppo e degli investimenti pubblici" funded by DPS (Dipartimento per lo Sviluppo e la Coesione economica)