MATH ONLINE TESTS: A RASCH ANALYSIS

Grazia Messineo¹, Salvatore Vassallo²

¹Università Cattolica Milano and IIS Falcone - Righi, Corsico (ITALY)
²Università Cattolica Milano (ITALY)

Abstract

The M.In.E.R.Va. project, which was presented at EDULEARN 2014 and 2016, was developed to help students of secondary schools to recover their deficiencies in Mathematics. As the experimentation has produced good results, we extended the project to Calculus and Financial Mathematics courses in Catholic University in Milan. We will use Rasch Analysis and other statistical techniques (e.g. Classical Test Theory and Item Response Theory) to prove the “goodness” of the multiple-choice questions that we used in the online exercises and to perform a differential analysis, dividing students in subgroups, in order to individuate the “disturbing” variables. We found that a few questions were problematic. In this case, the common approach is to change the question. We preferred instead to change the didactic approach to the subject of the question and to change the question only if the performance remains low. Since this trial showed quite good results, we can conclude that the platform can be successfully used to point out early problematic subjects and to intervene immediately.

Keywords: Online test, Rasch analysis, Item response theory. Classical test theory.

1 INTRODUCTION

In the last years there is an increasing need of powerful tools to analyse multiple choice test. Rasch models have been extensively used in research on measurements of learning, both in Italy and abroad. These methods began to spread when powerful computers started to be available for researchers to process complex statistical models. One of the main causes of the success of Rasch models is the opportunity to measure on a single scale both the ability of the involved subjects and the difficulty of the questions, so that it is immediately clear which questions have to be answered correctly by a student with a certain level of ability (according to the model) and which questions separate different levels of ability. This approach has some advantages with respect to the Classical Test Theory which is based on the percentage of correct answers.

Since 2011, M.In.E.R.Va project (Mathematics Interactive Exercises for Recovery and Enhancement) has been developed in Catholic University of Milan. This project has the aim of creating a tool to recover deficiencies in Mathematics for both Secondary School and University students. After various years, the need of verifying the adequacy of the tool has arisen, not only with a simple analysis of the improvements of the students, but also verifying that proposed exercises are suitable for the objectives of the project. Thus, we have begun to analyse some outcomes of the tests with Rasch model.

With the structure of the project is very easy to perform this kind of analysis, as the results are stored on a server in a csv file or a database, and thus they are immediately available for analysis. Moreover, tests are modular (each item is separated from the others) and so it is easy to remove questions that do not fit and put them in a different test.

This paper is divided in two parts: in the first one we will briefly describe Classical Test Theory, Item Response Theory and Rasch Model. In the second one we will analyse the data of one of the online math test given by students of Catholic University in two Academic Years. We will try to show that a change in the didactic approach for problematic items produces a better performance and becomes a valid tool for teachers to achieve a better understanding of the subject.

2 CLASSICAL TEST THEORY (CTT)

Classical test theory was proposed by Alfred Binet [1] to evaluate if the so-called test scores were valid and reliable instruments to measure psychological characteristics, which could not be measured directly. This unmeasurable characteristics are called latent variables.
Spearman ([2] and [3]) proposed in 1900 the fundamental equation of this theory, in which the total score of a person in a question of a test (denoted by $X_{ij}$ with $i$ is the number associated to the student and $j$ the number associated to the item) in a true score $\theta_i$, which measures the latent variable, and a random error component $\epsilon_{ij}$:

$$X_{ij} = \theta_i + \epsilon_{ij}$$

The basic assumptions of the model are:

1. True scores and measure errors are uncorrelated;
2. The expected measure error for all the subjects in the sample is zero;
3. Measure errors in parallel forms are uncorrelated.

We can thus affirm that

1. The expected value of observed scores is equal to the expected value of true scores;
2. The variance of observed scores is the sum of the variance of true scores and the variance of measure errors.

Reliability of measures is a very important characteristic for this model. Reliability can be defined as the stability of a measure for the same subject who answers to the same test in different moments (or to parallel forms of the test). The Cronbach alpha coefficient measures the internal consistency reliability, which can be thought as the expected correlation among scores of tests which measure the same latent variable:

$$\alpha = \frac{K}{K - 1} \left( 1 - \frac{\sum_{i=1}^{K} V(X_i)}{V(X)} \right)$$

CTT allows also to evaluate the quality of each item in the test. Two indexes can be used:

1. $p$-value or item difficulty: it is the percentage of subjects who answer correctly to the question;
2. Point-biserial correlation coefficient or item discrimination: it evaluates the capability of the item to discriminate correctly subjects with a high level of ability from students with a low level.

## 3 ITEM RESPONSE THEORY

Item Response Theory (IRT) tries to overcome the limits of CTT. Its aim is to build a model to explain each subject’s answers as a function of his ability and of the characteristics of the item itself. The probability of giving the correct answer to each question is thus a function of:

1. The person parameters, which measure the ability of each subject;
2. The item parameters, which measure the difficulty of the item, its discrimination power and the role of guessing in the choice of the correct answer.

A IRT model is defined by specifying its dimensional structure (in this models, unidimensionality is assumed) and the Item Characteristic Curve (ICC).

The Three parameter logistic model (3PL) is defined by the equation

$$P_j(\theta) = c_j + \frac{1 - c_j}{1 + e^{-D(\theta - \beta_j)}}, j = 1, 2, \ldots, m$$

where $P_j(\theta)$ is the probability of correct answer to the item $j$ as a function of the ability level $\theta$, $m$ is the total number of items in the test, $c_j$ is the guessing parameter (it justifies the fact that subjects with a low level of ability answer correctly to a difficult question), $\beta_j$ is the difficulty parameter, $\delta_j$ is the discrimination parameter and $D$ is a factor of scale.

In this general model

- If we consider $c_j = 0$ and $\delta_j = 1$, we obtain the One parameter logistic model (1PLM), in which only the difficulty parameter is supposed to determine the behaviour of the item.
• If we consider \( c_j = 0 \), we obtain the **Two parameters logistic model (2PLM)**, in which the behaviour of the items is determined by their difficulty and by their discrimination power among people who answer to the test (the parameter \( \delta_j \)).

• **3PL Models** consider also a third parameter, \( c_j \), which explains the so-called guessing, the fact that a subject answers correctly to a question he ignores.

We cannot say that models with a higher number of parameters are always better than models with a smaller one. As a matter of facts, parameters are not additive, so 2PLM and 3PLM do not guarantee **specific objectivity** (the property that guarantees that comparisons among students are item-independent and comparisons among items are students-independent) as the 1PLM does. Moreover, while estimating parameters, it is necessary to impose restrictions, in order to avoid non-acceptable values for discrimination and guessing parameters. For a wider explanation of these problems, see, for example, [4] and [5].

### 4 THE RASCH MODEL

Rasch Model can be thought as an IRT model (see [6], [7] and [8]), namely a 1PLM.

It is based on the assumption that a measure must show two characteristics:

- **Uni-dimensionality**: The latent variable measured must be one and only one, so each question must be designed to be a partial indicator of this variable. If the analysis shows that there is a secondary latent variable, the items measuring it must be separated from the others and put in a different test.

- **Specific objectivity**: Measures must not be influenced by individual characteristics other than the one we want to measure, by the other subjects tested or by the specific form of the test: Comparison between two people is **item-free** and the comparison between two items is **person-free**.

While IRT tries to find the best model to explain observed data, Rasch model is based on the assumption that the researcher must create a model satisfying the principle of specific objectivity. He/she must find out data which fit the model (see [8] and [9]). Rasch proposed in 1960 [10] a mathematical model which allows to compute the probability \( p_{ij} \) that the subject \( i \) with the ability \( \theta_i \) answers correctly to the question \( j \) which difficulty is measured by \( \beta_j \). So Rasch measure is based on the difference

\[
\theta_i - \beta_j
\]

where

\( \theta_i \) and \( \beta_j \) are, respectively, the score of \( i \)-th student, which indicates his/her ability, and the score of the \( j \)-th item, which indicates its difficulty.

For dichotomous data, the probabilistic model is:

\[
P(X_{ij} = 1|\theta_i, \beta_j) = \frac{e^{\theta_i - \beta_j}}{1 + e^{\theta_i - \beta_j}} = p_{ij}
\]

where \( X_{ij} \) is the result of person \( i \) on item \( j \), \( i = 1, \ldots, n \) and \( j = 1, \ldots, m \).

Data are collected in a raw score matrix, with \( n \) rows (one for each subject) and \( m \) columns (one for each item). Values can only be 0 or 1. The sum of each row \( n = \sum_{j=1}^{m} x_{ij} \) represents the total score of the subject \( i \), and the sum of each column \( c_j = \sum_{i=1}^{n} x_{ij} \) represents the score obtained by all the subjects in the item \( j \). The \( r_i \) statistic is sufficient for \( \theta_i \); The \( c_j \) statistic is sufficient for \( \beta_j \).

In order to measure ability and difficulty on the same scale, it is necessary to convert the values from the two measures into a common scale, which is the **logit**:

\[
\ln \frac{p_{ij}}{1 - p_{ij}}
\]

If we substitute equation (*) in equation (**), it is possible to define the parameters \( \theta_i \) and \( \beta_j \) in the same scale and to verify that the logit is equal to the difference \( \theta_i \) and \( \beta_j \). We can interpret this result in this way: The distance between the ability of a student and the difficulty of an item is equal to
logarithm of the odds ratio (the ratio of the probability of success and the probability of failure in the item).

5 ANALYSIS OF DATA OF THE M.IN.E.R.VA. PROJECT

In 2011 Catholic University of Milan granted a funding to a research project named M.In.E.R.Va. project, which provides a tool that allows students attending high schools to recover deficiencies in Mathematics. This project showed good results, so in 2014-15 Academic Year we extended it, as a training tool, to some groups of students of Catholic University, namely the ones who give a part of the final exam online. We presented this project and some of the results obtained at Edulearn 2014 and 2016 ([11] and [12]) and at ICETC 2014 [13].

Now we will analyse the data of one of the on-line tests given by the students of the course of Calculus in 2016 and in 2017. This test concerns Multivariate Calculus (which is a part of the exam program), it contains nine questions, on domains and on level lines of two variables functions. Rasch Model will be applied to data, to see if they fit the model and to analyse the goodness of items; distractor analysis will be performed with the techniques of CTT, to deeply investigate if some items need to be adjusted. We will also try to prove that if the teacher changes the didactic approach, following the results of the Rasch Analysis, students can achieve a better understanding of the subject and their performance on the test improves.

Analyses are performed both using some R packages for CTT, IRT and Rasch Model, and the software Winsteps®, specifically designed for Rasch Analysis.

5.1 Analysis of the results of the test given in 2016

In academic year 2015-16 the test on multivariate calculus was given by 28 students (this test is not compulsory, so students can decide to give it or not, to practice and to evaluate their performance on these matters).

5.1.1 Preliminary analysis

Fig. 1 shows the frequencies of total scores obtained by students in this test (each item is worth 1 point, except for item 3 which is worth 2 points).

![Figure 1. Distribution of total scores in the 2015-16 test.](image)

To perform the general analysis, data were converted to binary strings, with the correct answer coded 1 and the wrong answers coded 0. First of all, we must evaluate the internal consistency of items, i.e. their attitude to measure one single latent dimension. This characteristic is evaluated by the Cronbach alpha coefficient, an index proposed within CTT to evaluate internal consistency reliability. For our data, values of the Cronbach alpha index are shown in Table 1.
Table 1. Cronbach alpha for the 2015-16 data.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>0.3604</td>
<td>0.4328</td>
<td>0.1035</td>
<td>0.4293</td>
<td>0.3381</td>
<td>0.3785</td>
<td>0.1338</td>
<td>0.3452</td>
<td>0.3403</td>
<td>0.3789</td>
</tr>
</tbody>
</table>

Guidelines for interpretation of this index suggest that internal consistency is quite low (the index for all items is smaller than 0.6). Moreover, some of the items seem to contribute negatively to test consistency, as the value of the Cronbach alpha index is higher if we remove those items (the red values in Table 1).

Figure 2. Difficulty and discrimination indexes for the 2015-16 test.

Figure 2 shows the difficulty and discrimination indexes for the question in the 2015-16 test. Difficulty is the p-value of each item, computed as the percentage of total correct answers to the question. Questions are ordered from the most difficult (question 2) to the easiest (question 8). Discrimination index measures the capability of the question to discriminate between high and low performers. All the values of the two indexes are good, as they are all greater than 0.2. Discrimination index for item 8 could not be computed.

5.1.2 Rasch analysis

The preliminary results of the Rasch Analysis are shown in Figure 3.

Separation and reliability indexes are quite low: For students, values in the red box show Reliability = 0.29 < 0.8 and separation = 0.64 < 2; for items, values in the blue box show Reliability = 0.74 < 0.9 and separation = 1.70 < 3. This results tell us that the students’ sample and the number of items are not large enough to discriminate between high and low performers. There are no extreme items, i.e. items to which students gave all correct or wrong answers.
The brown box contains the values of the Rasch Measure, in logits. These values allow us to order questions, from the less difficult (items 7 and 8) to the most difficult (item 2).

Infit and outfit statistics measure how data fit the model. Outfit MNSQ values are acceptable if they belong to the range (0.5; 1.5); Z standard values are acceptable if they are in the range (-2;2). We can see that all values are acceptable, except of the one of the eighth question (the same for which we could not evaluate discrimination index, see par. 5.1.1).

Violet box contains the values of the point-measure correlation, which allow us to evaluate the correlation between observed scores and Rasch Measures for each question. These values are quite low, and there is, once again, a problem with the eighth question, which has a negative value of PTMEA.

Figure 4. Person-Item map for 2015-16 test.

Figure 4 shows the Person-Item map for the 2015-16 test. Here students' ability and items difficulty are represented on the same scale, students in the higher part and items in the lowest one. Students with a certain level of ability have a probability of 50% to answer correctly to a question with the same level of difficulty, a probability higher than 50% to answer correctly to a question with a level of difficulty lower than their ability and a lower probability in the reverse case. For this group, we can see that there are some students whose level of ability is not high enough to allow them to answer correctly to any of the questions, and that there are some other ones whose level of ability should allow them to answer correctly to all the questions.

Figure 5 shows a part of Winsteps® table 26.3, in which items are analysed in detail and problematic ones can be detected. Items are shown once again in correlation order, from the less correlated to the most correlated to the total score. We can notice that item 8 (together with 3 and 9) is problematic: The asterisk placed next to the measure associated to the correct answer tells us that the ability of the students does not increase if they respond correctly. Students’ measure is expected to be the highest for that item if a student gives the correct answer and the lowest if he gives the wrong ones.

Figure 5. Distractor Analysis - item 8

Figure 6. Distractor Analysis - item 1.
Figure 6 shows item 1, an example of well-behaved item: Students’ ability associated to the correct answer is the highest one, and also because only the correct answer is positively correlated to the test total score. Item 8, instead, shows problems both in students’ ability (students who answer correctly have an average ability lower than students who give, for example, the wrong answer e). Moreover, not only the correct answers are positively correlated to the test score, but also almost all the wrong answers (and wrong answer d has a higher correlation than the correct one).

Figure 7 shows the distractor analysis for item 1 (the well-behaved one): The percentage of choices of the correct option (red line) increases with students’ ability and it is always greater than the percentage of choices of the wrong answers. Figure 8 shows the same data for item 8 (the wrong-behaved one): The percentage of choices of the correct option does not increase with students’ ability and the percentage of choices of one of the wrong answers increases with it.

5.2 Analysis of the results of the test given in 2017

As the analysis of data of 2015-16 showed that some subjects were not well understood, in the next academic year we decided to change the didactic approach to them, in order to establish if students’ performance on the problematic questions would improve or not.

5.2.1 Preliminary analysis

In 2016-17 the same test was given by 29 students.

Figure 9 shows the distribution of total scores obtained. Values of the Cronbach alpha coefficient are shown in Table 2.
### Table 2. Cronbach alpha for the 2016-17 test.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>0.7738</td>
<td>0.7645</td>
<td>0.7060</td>
<td>0.7625</td>
<td>0.7240</td>
<td>0.7489</td>
<td>0.7046</td>
<td>0.7502</td>
<td><strong>0.8045</strong></td>
<td><strong>0.7843</strong></td>
</tr>
</tbody>
</table>

Guidelines for interpretation of this index suggest that internal consistency is quite good (the index for all items is between 0.7 and 0.8). Also this time, some of the items seem to contribute negatively to test consistency, as the value of the Cronbach alpha index is higher if we remove those items (the red values in Table 2).

![Figure 10. Difficulty and discrimination indexes for 2016-17 test.](image)

Figure 10 shows the difficulty and discrimination indexes for the question in the 2016-17 test. Questions are ordered from the most difficult (question 2) to the easiest (question 7 for these students). All the values of the two indexes are good, as they are all greater than 0.2. Discrimination index for item 8 could not be computed also this year.

#### 5.2.2 Rasch analysis

The preliminary results of the Rasch Analysis are shown in Figure 11.

Separation and reliability indexes are quite low, although better than the 2015-16 ones: For students, values in the red box show Reliability = 0.44 < 0.8 and separation = 0.88 < 2; for items, values in the blue box show Reliability = 0.79 < 0.9 and separation = 1.96 < 3. This results tell us once again that the students' sample and the number of items are not large enough to discriminate between high and low performers. There are no extreme items, i.e. items to which students gave all correct or wrong answers.

The brown box contains the values of the Rasch Measure, in logits. These values allow us to order questions, from the less difficult (item 7) to the most difficult (item 2).

Outfit MNSQ values are all acceptable.

Violet box contains the values of the point-measure correlation, which which are this year all better than the previous ones. Moreover, no question has a negative correlation with the test total score. All these indexes allow us to conclude that, from a general point of view, students of this year performed better than students of the previous year: it seems that the different didactic approach has worked as expected.
Figure 11. Analysis of results of the 2016-17 test.

Figure 12. Person-Item Map for the 2016-17 test. We can notice that for this group, more students than in the previous year have a level of ability higher than the difficulty of all items, who are expected to answer correctly to all the questions proposed. There are also more students with a level of ability lower than the difficulty of all the questions, so we can expect these students to give the wrong answer to all the items.

Figure 13. Item 8 in the 2016-17 test

Figure 12 shows the Person-Item Map for the 2016-17 test. We can notice that for this group, more students than in the previous year have a level of ability higher than the difficulty of all items, who are expected to answer correctly to all the questions proposed. There are also more students with a level of ability lower than the difficulty of all the questions, so we can expect these students to give the wrong answer to all the items.

Figure 13 shows part of Winsteps® table 26.3, in which we can notice that now item 8 (together with item 2) is problematic: the asterisk placed next to the measure associated to the correct answer tells us that the ability of the students does not increase if they respond correctly. Moreover, once again some of the wrong answers are positively correlated with the test total score. In addition, we have also some questions (see for example question 3 and 7) in which only some distractors were chosen. This behaviour is not what we search for, because we need distractors which are sufficiently attractive for students with a low level of ability to be chosen at least once and to contribute to the discrimination power of the item. For this group, distractors for items 3 and 7 are not sufficiently attractive.
Figure 14 shows the distractor analysis for item 8 in the 2016-17 test: We can notice a better behaviour of the item, as the probability of correct answer increases for the group of best performers, as the probability of wrong answer decreases. The behaviour of the distractors is not, once again, good for students with an intermediate level of ability.

6 CONCLUSIONS

Rasch Analysis has allowed us to give a precise evaluation of the abilities of the students who participated in the M.In.E.R.Va. project and, at the same time, to evaluate the effectiveness of the exercises proposed. This allowed us to promptly modify the questions proposed or the structure of the test. This is particularly simple in the M.In.E.R.Va. project, because of its modularity.

We have also seen that, to improve students’ performance, it may be useful to change the didactic approach following the suggestions of a statistical analysis similar to the one performed, which gives precise indications on the problematic areas and on the most frequent errors.

REFERENCES
