MODESTUM
OPEN ACCESS

EURASIA Journal of Mathematics, Science and Technology Education, 2018, 14(7), 2895-2906
ISSN:1305-8223 (online) 1305-8215 (print)
Research Paper https://doi.org/10.29333/ejmste/90988

The Didactic Contract to Interpret Some Statistical Evidence in Mathematics Standardized Assessment Tests

Federica Ferretti [1], Chiara Giberti [2*], Alice Lemmo [3]

[1] Free University of Bolzano-Bozen, Bolzano, ITALY
[2] University of Bologna, Bologna, ITALY
[3] I.C. Bassa Anaunia- Duenno, Trento, ITALY

Received 12 January 2018 ▪ Revised 18 March 2018 ▪ Accepted 6 April 2018

ABSTRACT

In this study we analyse results of Italian standardized tests in mathematics integrating quantitative analysis based on the Rasch Model and didactical interpretation. We use specific graphs to analyse the trend of each answer as function of the students' math ability. This approach led us to focus on specific items in which a wrong answer results particularly popular among medium/high level students and analyse this particular trend with the lenses of math education theories. The study reveals that these phenomena are particularly related to implicit and explicit rules governing classroom practices exist at all school levels and regard different mathematical content and skills.

Keywords: didactic contract, mathematics education, mixed methods, Rasch analysis, standardized assessment

INTRODUCTION

In recent years, international standardized assessment such as OECD-PISA and TIMMS tests gained an increasing interest in particular concerning public opinion. At the same time, the results of this tests can be extremely productive in math education research field (Leder & Lubienski, 2015; Middleton, Cai & Hwang, 2015) but their usage is growing only in the last few years (Sfard, 2005).

Every year in Italy, the National Evaluation System entrust the administration of large scale tests to INVALSI[1]. The data collected and the analysis of results by INVALSI highlight macro-scale phenomena. Specifically, regarding the standardized mathematics tests, students' answers to some of the test items reveal behavioural attitudes that allow us to understand more in depth peculiarities of teaching and learning process and some causes of difficulty nationwide.

Recently, an increasing number of research studies have focused on students' math performances facing on national and international standardized assessment (Anderson, Lin, Treagust, Ross, & Yore, 2007; Arzarello, Garuti, & Ricci, 2015; Di Tommaso, Mendolia & Contini, 2016; Middleton et al., 2015); in details, some of these (i.e. Bolondi et al., 2016; Bolondi, Cascella, & Giberti, 2017; Mellone, Romano, Tortora, Statale, Mangino, & Pagani, 2013) show how often the didactic phenomena that emerged through statistical methods are not directly related to the mathematical content, but can be interpreted with mathematics education constructs, such as the didactic contract. In this research we went further that and we looked for statistical evidence that would allow us to have further information.

This research sets out to interpret and analyse some of the findings which emerge across the country in statistical analysis of the INVALSI tests, and which regard all school levels involved in the tests. For each test considered in this paper, the data analysed are those of INVALSI's sample which consists of approximately 30,000 students and

[1] From the official website of the INVALSI: INVALSI is a research institute with the status of legal entity governed by public law. The Institute carries out periodic and systematic checks on students' knowledge and skills, and on the overall quality of the educational offering of schools and vocational training institutes, also with a view to lifelong learning; in particular, it runs the National Evaluation System.

✉ federica.ferretti@unibz.it ✉ chiara.gib@gmail.com **(*Correspondence)** ✉ alice.lemmo@gmail.com

Contribution of this paper to the literature

- The analysis of didactic phenomena emerged from standardized assessments fits in the international research in mathematics education and the introduction of innovative investigation methods, such as that shown in this research, are relevant for the research community.
- The innovative method of this research constitutes a development of existing research in mathematics teaching using data from standardized surveys as:
  - combines qualitative methods with quantitative methods extrapolating information on students' answers from statistical elaborations
  - consists in an item-level analysis that allow to understand deeper causes of specific students' answers and interpret them with the lenses of mathematics education theories
  - in particular, allows us to confirm and, therefore, study and develop the educational contract construct
  - the results obtained provide information for a well-defined category of students, identified on the bases of their mathematical ability level measured on the entire test.

it is representative of the population of Italian students attending that grade. The INVALSI statistical analyses are based on Classical Test Theory to provide the consistency of the test and on Item Response Theory (IRT) models, and in particular the Rasch Model to analyse items' features (INVALSI, 2017).

The Rasch Model (Barbaranelli & Natali, 2005; Rasch, 1960) is a one-parameter logistic model, it belongs to the *Item Response Theory* (IRT) category and produces a jointly estimate of the difficulty parameters of each test item and the ability parameter of each student.

Using this model it is possible to push the analysis of standardized assessment, often focused on the entire tests' results, to an item-level. As highlighted by Leder and Lubienski (2015), item level analysis of standardized assessment is particularly important because gives information concerning specific difficulties of students in mathematics and this allow researchers and teachers to intervene effectively.

In particular, item-level analysis provided by the Rasch Model allows us to express the probability of supplying the correct response to a test item on the basis of the difficulty of the question itself and the ability of the student as evaluated via the entire test. In general, higher student ability is matched by increased probability that he/she will produce the correct answer, whilst the number of wrong answers should decrease as student ability level rises. However, there are cases where this trend is not displayed – some items may offer a wrong answer which results particularly popular with medium/high level students.

Understanding the cause of this phenomenon is complex as various educational factors come into play, as has been frequently highlighted in the national and international literature of mathematics education. It concerns issues linked to implicit and explicit rules which govern mathematical practices in classrooms, particularly regarding the solving of mathematics tasks, which regulate the certainties and behaviour of students.

The study reveals that these phenomena exist at all school levels and regard different mathematical content and skills: a detailed key to such behaviour involves some well-known mathematics education constructs.

Moreover, this paper presents a research on the INVALSI mathematics test questions, beginning with the quantitative data collected by the National Evaluation System. The aim of the study is to propose an integrated analysis of the tasks which allows us to interpret some of the phenomena emerging in the quantitative results via qualitative analysis.

The introduction, both nationally and internationally, of standardized assessment tests such as OCSE-PISA, IEA-TIMSS, and INVALSI, can provide important data at a systemic level for mathematics education research (Looney, 2011). Although the main aim of this analysis is an assessment of the education system, and abilities and skills achieved by students at different scholastic levels, analysis of task wording and student performance may also offer important data for research purposes.

From this point of view, one "critical issue, which remains partially unresolved, regards the 'translation' of the quantitative statistic results of the national sample into information and proposals that may become effective driving forces of innovation rather than pure data which leave the door open to interpretations (often hasty and inadequately considered) that end up deeply distorting the objectives of the evaluation itself" (Bolondi et al., 2016). As explained before, to move in this direction, it is useful an item-level analysis that points out specific difficulties/strength of students and this can be supported by a qualitative analysis of the most interesting items throughout the lenses of math education theories. However, in some studies the test results have revealed some very interesting macro-scale phenomena (Branchetti et al., 2015; Bolondi et al., 2016; Bolondi, Cascella, & Giberti, 2017), such as new effects of the didactic contract; these findings may then be studied in depth via a mixed-method approach from QUAN to QUAL, performing first a quantitative analysis and later a qualitative analysis (Johnson & Onwuegbuzie, 2004).

The main aim of this study is to highlight the potentialities of an item-level analysis of standardize test data, combining both quantitative and qualitative analysis. Our hypothesis is that item-level analyses may provide a huge amount of information refer to teaching and learning processes.

THEORETICAL AND METHODOLOGICAL FRAMEWORK

Although Classic Test Theory (CTT) offers important statistical tools for the assessment of tests (Barbaranelli & Natali, 2005; INVALSI, 2017), the analytical studies presented in this paper and in reports of INVALSI test results are mainly based on the more modern Item Response Theory. This latter solution makes use of various mathematical models to measure latent variables and allows us to overcome the principal limitations of CTT, such as the dependence between estimated student ability and item difficulty.

In this context, we will consider the simplest IRT model which is also used in the main analytical studies of INVALSI tests: the Rasch model (INVALSI, 2017; Rasch, 1960).

The Rasch model is a one-parameter logistic model, and thus the simplest of the IRT models. It allows us to calculate the probability of correct response to a determined item, according to the ability of the student and the psychometric characteristics of the item itself (particularly, the item's difficulty).

The relationship between the student's ability and the probability of correct response to the item may be represented in a graph via a theoretical curve known as the Item Characteristic Curve (ICC).

Once the Rasch model is applied, we have the possibility to move our research to item-level analysis by observing features of specific graph output of the model. The examination of these graphs, called *distractor plots,* makes it possible to extract extremely useful information regarding each individual item. In the same graph as that where the ICC is tracked regarding correct answers (usually a continuous blue line – **Figure 1**), it is also possible to view empirical data regarding the correct answer and other alternative responses. In this way, it is possible to observe to what extent the empirical curve of correct answers be coherent with the theoretical curve, and also to analyse the trend of each distractor (i.e. incorrect response) according to the students' level of ability.

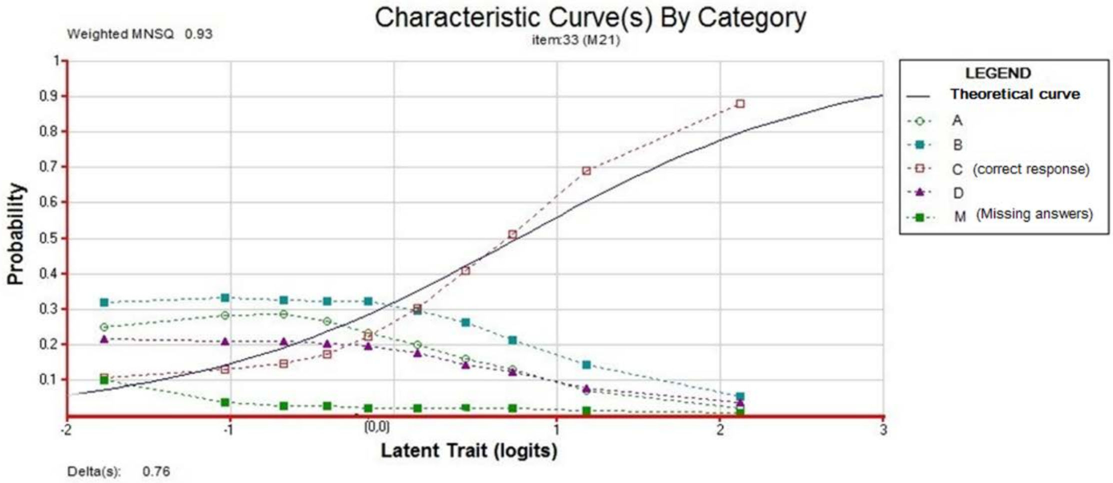


Figure 1. Example of a distractor plot for one item

The distractor plot on the x-axis (**Figure 1**) reports the Rasch score in terms of student ability across the entire test and, as already outlined, the continuous line represents the model's theoretical curve (ICC) which reveals the probability of correctly answering the questions according to student ability. The broken lines represent empirical data collected on each reply option for the item used. For the graph implementation, students are divided into deciles according to their level of ability as measured by the entire test, and, for each group of ten, the percentage of students who chose each answer is reported.

In this item, the comparison between the empirical correct answer trend and the theoretical curve results as acceptable (weighted = 0.93); it also appears, however, that in this case the model tends to overestimate the students with medium-low ability levels whilst underestimating high-level students. The most attractive option is B, which was chosen by a high percentage of students, including those of medium and medium-high level abilities. The other two options also display "good functioning" and were chosen by students of low and medium ability. Finally, it can be noted that only a few students did not answer the question, almost all of whom belonged to the decile with the lowest ability level.

The Rasch model can only be applied in cases where certain conditions exist, allowing the application of the model and estimate of parameters (Barbaranelli & Natali, 2005; Hambleton, Swaminathan & Rogers, 1991). In particular, unidimensionality, local independence and monotonicity of the test are required. The condition of monotonicity demands that (for each item) the probability of a correct response grows monotonically with the increase of the students' ability level and can be checked via the graphic representation of empirical data, i.e. by tracking the distractor plots.

From a strictly statistical point of view, it could be expected then that a higher level of student ability correlates with a higher percentage of correct answers for an item and, simultaneously, a lower percentage of wrong answers. Observing the responses to multiple choice questions (where there is only one correct answer but another two/three wrong answers are also suggested), it can be seen that the percentage of wrong answers given (considered as a whole and bearing in mind the missing answers) always decrease with the students' ability but, if the relative curves of the individual distractor items are considered separately, it is possible to see answers' trends which are not strictly decreasing (for example, the curve for option B in Item 24: 2, **Figure 2**).

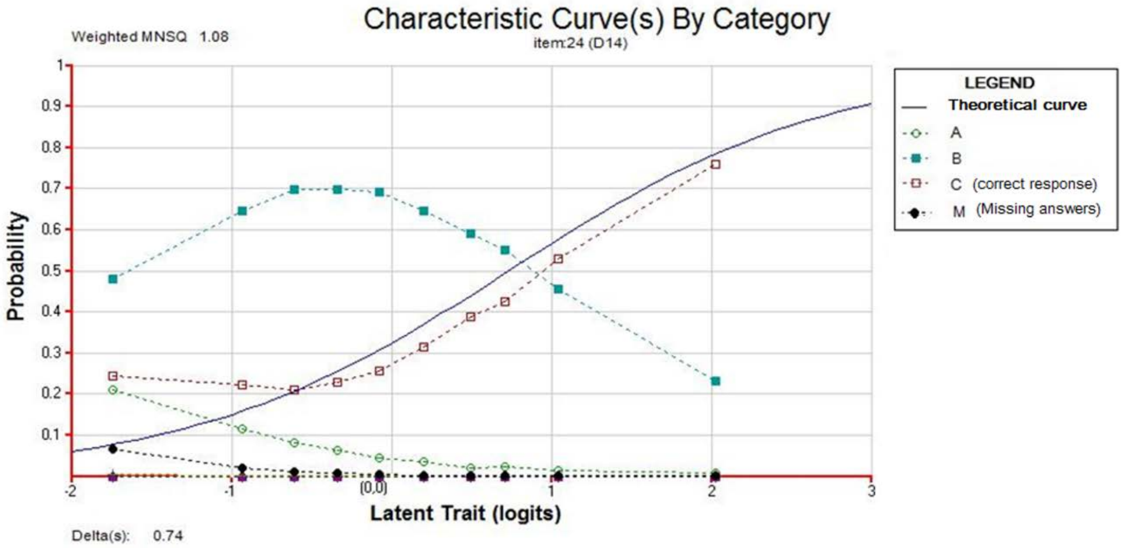


Figure 2. Example of a distractor plot for an item with decreasing performance in a curve regarding a wrong option

In the example shown in **Figure 2**, the curve related to option B reveals an increasing performance followed by a decrease which will henceforth be indicated as a "humped performance": the percentage of students choosing this option increases as student ability level rises in the low and medium-low deciles, whilst only from the fifth decile onwards does the curve show a decrease.

Analysis of this phenomenon (the "humped performance" of an option) is complex as various interactive factors come into play: students with varying levels of ability may encounter different obstacles when faced with a task, supply wrong answers for different reasons, and favour one wrong answer over another as a result of different approaches and problems.

Possible causes for the selection of wrong answers are often linked to factors regarding implicit and explicit rules established during teaching/learning processes which regulate mathematical task activity in class and often lead to wrong convictions. For an in-depth understanding of the reasons behind such circumstances, it is necessary to carry out a critical analysis of responses to the individual tasks via some mathematical education notions.

In this research, we focus on items from various school levels (from primary to high school) which display good measurement properties (Barbaranelli & Natali, 2005; INVALSI, 2017) and in which at least one option of response demonstrates a "humped performance" that may be linked to teaching factors. In particular, in the following examples, one of the main constructs that can supply a key to reading statistical results of this type at a systemic level is the *didactic contract*.

The didactic contract forms part of Guy Brousseau's Theory of Didactical Situations in Mathematics and refers to the set of the teacher's behaviours as expected by the student, and the set of student's behaviours as expected by the teacher (Brousseau, 1988; EMS-EC, 2012).

Specifically, in a teaching situation, prepared and carried out by a teacher, the student is generally given the task of resolving a (mathematical) problem, but the key to this task is found by interpreting the given questions, supplied responses, and the obligations imposed by the teacher's methodology. These (specific) habits of the teacher as expected by the student and the student behaviour expected by the teacher form the didactic contract

(Brousseau, 1980). This notion supplies keys to interpreting the various situations that emerge during classroom activity and has revealed itself to be a particularly useful tool in interpreting situations that arise during mathematical task-solving, also in standardized test conditions (Bolondi et al., 2016; Ferretti, 2015).

As will be revealed later, some facts which emerge can be analysed with notions already mentioned in the literature, such as the clause of the didactic contract entitled the "need for formal justification", and other ad hoc constructs such as the "Age of the Earth" effect of the didactic contract (Ferretti, 2015).

QUALITATIVE AND QUANTITATIVE ANALYSIS OF QUESTIONS: SOME EXAMPLES

The study presented in this article began by selecting INVALSI items which displayed a "humped performance" for at least one of the option items.

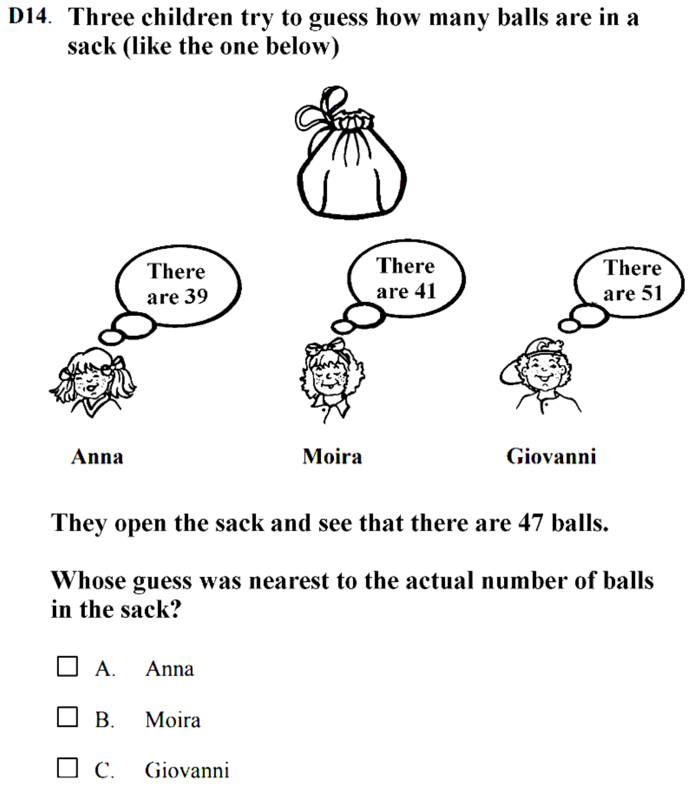
From an initial qualitative analysis of the individual item, it was revealed that in the majority of cases the response options included one linked to difficulty as highlighted in mathematical education research, and this precise option was linked to a "humped performance" by one of the distractors.

In the following section we present some examples of analysis of items referring to different mathematical fields and school levels.

The items selected are from INVALSI tests administered in different years (from 2011 to 2017) and at different school levels (from grade 2 to grade 10). The statistical analyses presented conform to those adopted by the INVALSI Institute during test validation and analysis of results, and are based on INVALSI samples comprising (for each test) 30,000 – 40,000 students nationwide. All the selected questions are multiple choice options and reveal good functionality in terms of model fit (weighted), discrimination, percentage of correct answers and point-biserial correlation between each answer and the overall ability of students (a negative correlation for wrong answers, and positive for correct answers).

In carrying out the study, we used the research tool GESTINV database (Gestinv 2.0., 2018; www.gestinv.it) which has already been verified in other research studies (Ferretti & Gambini, 2017) and from which we extracted all the results and graphs presented below.

The first example reported (Figure 2) refers to question D14 administered in the grade 2 mathematics INVALSI test of 2010/2011 (Figure 3), which belongs to the content area: "Numbers".



D14. Three children try to guess how many balls are in a sack (like the one below)

They open the sack and see that there are 47 balls.

Whose guess was nearest to the actual number of balls in the sack?

□ A. Anna

□ B. Moira

□ C. Giovanni

Figure 3. Question D14 from the grade 2 test of 2011

From a quantitative point of view, it can be seen that this is quite a difficult item (delta=0.74) to which only 35% of students answered correctly. **Table 1** reveals that there is a very low percentage of missing answers (less than 2%), and most of the students who made a mistake chose option B (57% of the total answers).

Table 1. IRT and percentage of responses of item D14 from the grade 2 test of 2011

ITEM D14 – STATISTICAL FEATURES			
Cases for this item 31842 **Item Threshold(s): 0.73**	**Discrimination 0.29** **Weighted MNSQ 1.08**	**Item Delta(s): 0.74**	
Label	**Score**	**Count**	**% of tot**
A	0.00	2087	6.55
B	0.00	18216	57.21
C (correct)	**1.00**	**11104**	**34.87**
Missing	0.00	435	1.37

Furthermore, again from the data reported in **Table 1** and looking at the distractor plot (**Figure 2**), it can be noted that the item has a good functionality from a psychometric point of view: its fit with the model data is acceptable (weighted = 1.08) and it discriminates well between respondents with high and low levels of ability (discrimination = 0.29).

The item asks students to compare three natural numbers and identify the closest number to another one given. It is a question designed to gauge the students' ability in estimating and comparing natural numbers. The correct response is C; the other two response options are linked to two problem areas. Specifically, those students who chose option A may have concentrated only on the figure representing the tens of the number 47.

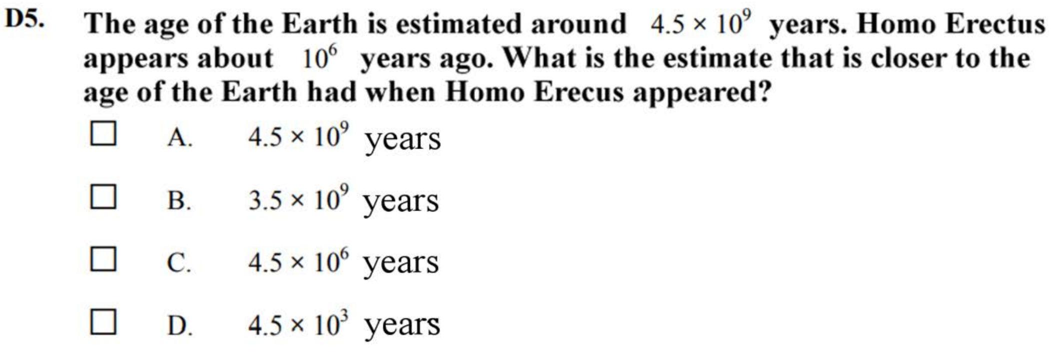
From this perspective, the students may have identified the number 39 as that closest to 40, revealing a difficulty in ordering natural numbers. The option A curve is decreasing and shows that the students who chose this option (around 6%) belong to the group of students displaying a low ability level in the test.

The option B curve on the other hand reveals a "humped performance" and is particularly attractive to students of medium level ability; furthermore, this option was chosen by a large percentage of students also among the higher level groups, and only in the last two deciles was the percentage of correct answers higher than that of distractor B. One possible reason for so many "good" students having chosen option B (41) as the answer could be linked to the fact that, in effect, this number is the only one proposed with a place value in tens the same as that of the number of balls in the sack (47). The students who chose this response may have been influenced by this similarity in tens, without considering the need to identify the "nearest" natural number.

In this case, then, it may be supposed that the students have a partial awareness of place-value notation of a number and thus stop only to consider the tens of the figure without comparing the entire number. The percentage of students who chose this option was 50% and an analysis of the distractor plot reveals that this did not only comprise students of low ability in test performance but also students with medium level scores. This fact reveals an interesting characteristic, i.e. that most students who showed difficulty in estimating and ordering natural numbers are mainly those who achieved average performance levels in the test.

One significant issue lies in the fact that in didactic practices usually the concept of estimation is tackled in the sense of "approaching something", implicitly meaning a "rounded-down estimate". The classroom task habits are revealed also in student performance during mathematical tasks, as a consequence of the didactic contract (Brousseau, 1988); the "humped performance" of the option which presents a rounded-down estimate may confirm the influence of the didactic contract in students' choices.

Another item of interest is question D5 from the grade 10 mathematics INVALSI test of 2011 (**Figure 4**).



D5. **The age of the Earth is estimated around 4.5×10^9 years. Homo Erectus appears about 10^6 years ago. What is the estimate that is closer to the age of the Earth had when Homo Erecus appeared?**

□ A. 4.5×10^9 years

□ B. 3.5×10^9 years

□ C. 4.5×10^6 years

□ D. 4.5×10^3 years

Figure 4. Question D5 from the grade 10 mathematics test of 2011

As we can see in **Table 2**, the correct answer (A) is chosen by only slightly more than 10% of students.

Table 2. IRT and percentage of responses of item D5 from the grade 10 maths test of 2011

ITEM D5 – STATISTICAL FEATURES			
Cases for this item 43458 **Item Threshold(s): 2.55**	**Discrimination 0.32** **Weighted MNSQ 0.97**	**Item Delta(s): 2.56**	
Label	**Score**	**Count**	**% of tot**
A (correct)	**1.00**	**4438**	**10.21**
B	0.00	2992	6.88
C	0.00	10084	23.20
D	0.00	24831	57.14
Missing	0.00	1113	2.56

The low number of correct responses highlights the difficulties students had in carrying out approximate estimation, numerical estimates, and ordering of numbers, as already revealed in the previous levels. One significant finding is that the correct response is one of the explicit data presented in the text and the failure to choose the correct response is part of a wider phenomenon already analysed in studies by Ferretti (2015) which reveal a new effect of the didactic contract – the "Age of the Earth" effect. Option B could have been chosen by students who mistakenly subtracted 10^9 from the age of the Earth without considering the difference in size ordering, whilst option C may have been chosen by students who made a mistake in the subtraction itself. In observing the Distractor Plot (**Figure 8**), it can be seen that such choices were mostly made by students who performed weakly in the test. In fact, the graph data reporting the choice of the two options decreases in line with higher student ability.

The most interesting option in this case is D, the most popular choice (selected by almost 60% of the sample group). This option may have been chosen by students who subtracted the exponent 6 present in the figure referring to the estimated time of arrival on Earth of Homo Sapiens from the exponent 9 in the estimated age of the Earth. From this point of view, the students may have mnemonically applied calculus tables linked to the property of powers.

The first interpretations of the phenomenon highlighted in the Age of the Earth question have linked student behaviour generically to the effects of the didactic contract as outlined by Brousseau (D'Amore, 2008). When given two numbers to the power of x, performing the subtraction of the exponents to carry out the subtraction of the numbers themselves represents a familiar operation to students as regards content but something which is completely wrong from a mathematical point of view. This behaviour can be tracked to a well-documented feature of the didactic contract, the need for formal justification (D'Amore, 2008).

As we can see from the Distractor Plot (**Figure 5**), the option D is the one favoured at all levels of ability, and the most popular choice of option for students of medium-high ability in the latent character scale; once again the curve relating to this option reveals a "humped" effect.

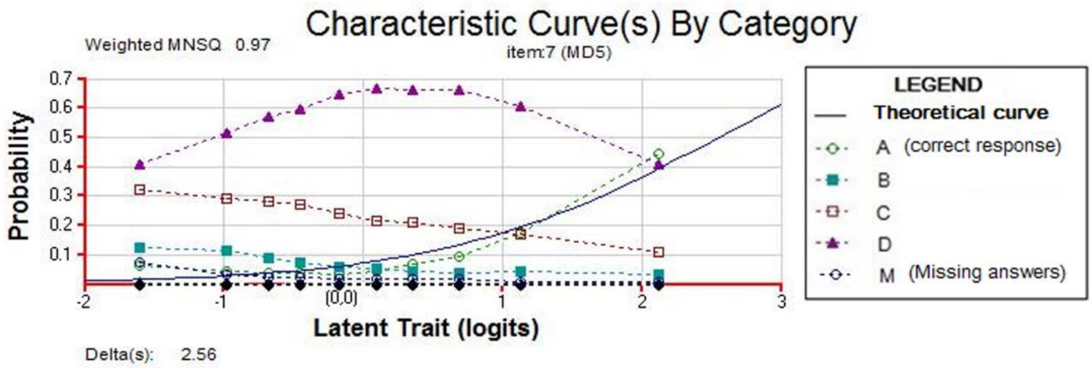
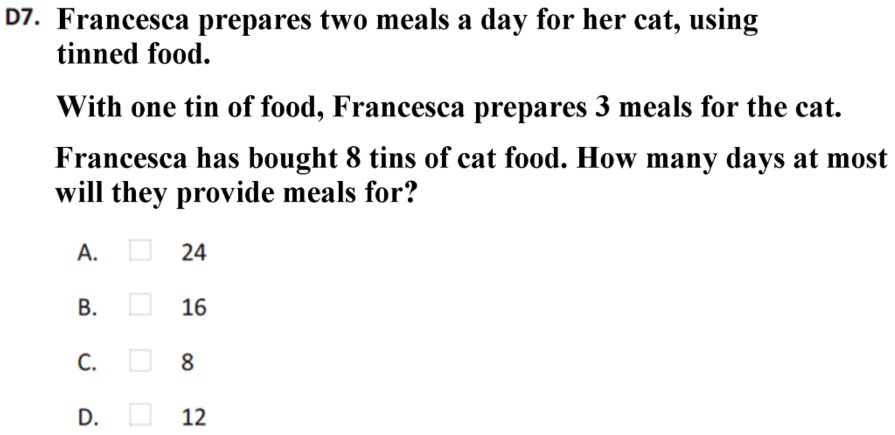


Figure 5. Distractor plot of item D5 from the level 10 mathematics test of 2011

The difficulties described above refer to the same mathematical field (Numbers) although at different scholastic levels. However, it is possible to identify questions with similar trends in other fields. The next example presents a task regarding the content "Relations and Functions" administered in the grade 5 mathematics INVALSI test of 2015 (**Figure 6**).



D7. Francesca prepares two meals a day for her cat, using tinned food.

With one tin of food, Francesca prepares 3 meals for the cat.

Francesca has bought 8 tins of cat food. How many days at most will they provide meals for?

A. □ 24

B. □ 16

C. □ 8

D. □ 12

Figure 6. Item D7 of the grade 5 mathematics test of 2017

This is a multiple choice item with the correct response being option D, which was chosen by fewer than 30% of students. To resolve this problem, the student must consider the entire text, understand the situation outlined, and focus not only on the numeric data give but also on the written textual content. As can be seen in **Table 3**, the question was quite difficult (delta = 1.10) and operates well in terms of fit with model (weighted = 1.04) and discrimination (discrimination = 0.35).

Table 3. IRT and percentage of responses of item 7 from grade 5 test of 2015

ITEM D7 – STATISTICAL FEATURES			
Cases for this item 22030	**Discrimination 0.30**	**Item-Total Cor. 0.35**	
Item Threshold(s): 1.10	**Weighted MNSQ 1.04**	**Item Delta(s): 1.10**	
Label	**Score**	**Count**	**% of tot**
A	0.00	10058	45.66
B	0.00	2407	10.93
C	0.00	2969	13.48
D (correct)	**1.00**	**6363**	**28.88**
Missing	0.00	255	1.16

Option B may have been chosen by students multiplying the number of tins by the number of meals per day. Option C could have been the choice of students who focused only on the number of tins without considering the information regarding the number of meals per day that Francesca prepares for her cat.

Regarding the Distractor Plot (**Figure 7**), it may be noted that these two options operate in a classic manner as distractors: both display decreasing monotonic function. The curve representing option A, on the other hand, shows a totally different behaviour from the other distractors and results as an option particularly attractive to respondents of medium ability. This option was chosen by a high percentage of students of every level of ability: in the lowest ability decile, it was chosen by almost 40% of students and only the two highest ability deciles favoured the correct response over this distractor.

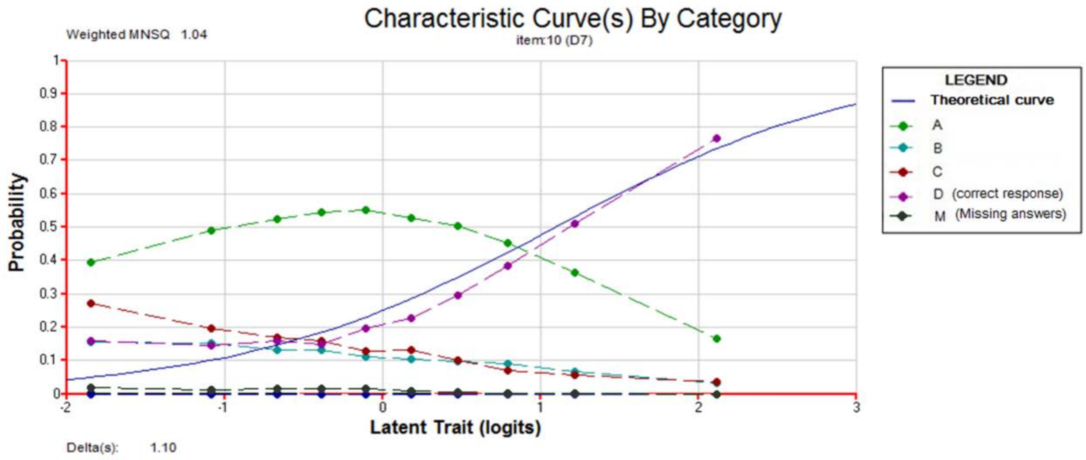
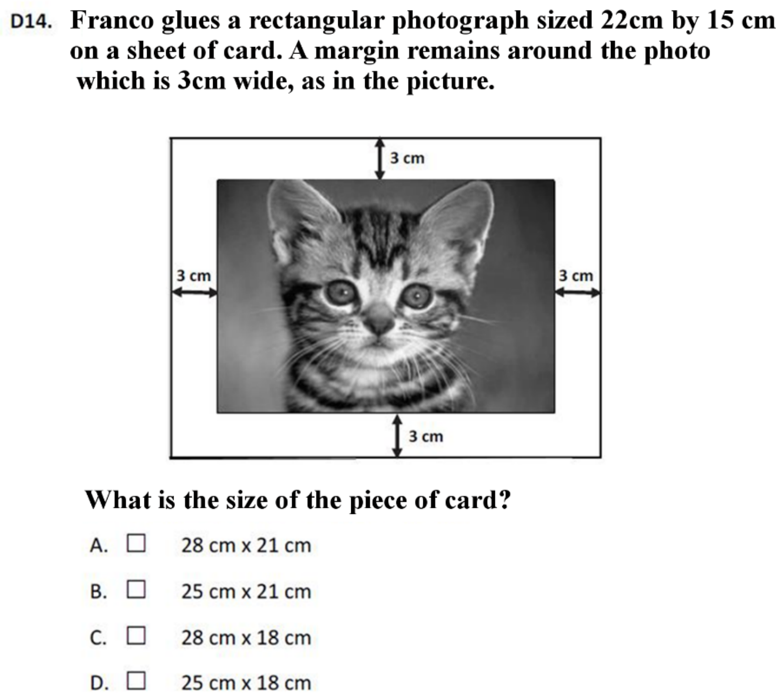


Figure 7. Distractor plot for item 7 from the grade 5 mathematics test of 2015

However, it can be seen that the choice of option A was highest amongst students of medium-level ability – more than 50% of students between the third and seventh decile groups chose this response option. This can be explained by referring to the "need for formal justification" hypothesis (D'Amore, 2008); it is possible that the students who chose this option did indeed identify all three data items present in the text but failed to grasp the problem situation posed and instead multiplied the figures in the text without checking the appropriateness of the calculation with regard to the situation presented. This type of behaviour is probably due to a teaching methodology based mainly on procedure; students who must resolve a problem tend to be asked to focus their attention on identifying data and the operation to be performed without reflecting more deeply on the situational problem posed. As you can see in the graph (**Figure 7**), students most affected by such a methodology are those of medium-level ability, who manage to identify the data presented in the question but in order to work out the solution turn to a procedure, to the identification of an operation that may however cause a loss of intended meaning and the wrong contextualization of the result.

Analogous behaviour can be noted in the content "Space and Shape" item; for example, in item D14 of the grade 6 mathematics test of 2013 (**Figure 8**).


3 cm
3 cm
3 cm
3 cm

D14. **Franco glues a rectangular photograph sized 22cm by 15 cm on a sheet of card. A margin remains around the photo which is 3cm wide, as in the picture.**

What is the size of the piece of card?

A. □ 28 cm x 21 cm

B. □ 25 cm x 21 cm

C. □ 28 cm x 18 cm

D. □ 25 cm x 18 cm

Figure 8. Item D14 from the grade 6 mathematics test of 2013

This case also comprises a multiple choice item with four response options, only one of which is correct (option A).

Table 4 breaks down the percentage of responses; only 26% of students replied correctly to the task. Option D, which was the most popular (chosen by around 50% of students) shows a "humped performance" in the distractor plot (Figure 9).

Table 4. IRT and percentage of responses of item D14 from the grade 6 test of 2013

ITEM D14 – STATISTICAL FEATURES			
Cases for this item 27416 **Item Threshold(s): 1.17**	**Discrimination 0.29** **Weighted MNSQ 1.06**	**Item Delta(s): 1.17**	
Label	**Score**	**Count**	**% of tot**
A (correct)	**1.00**	**2747**	**26.43**
B	0.00	2514	9.17
C	0.00	2591	9.45
D	0.00	14235	51.92
Missing	0.00	829	3.02

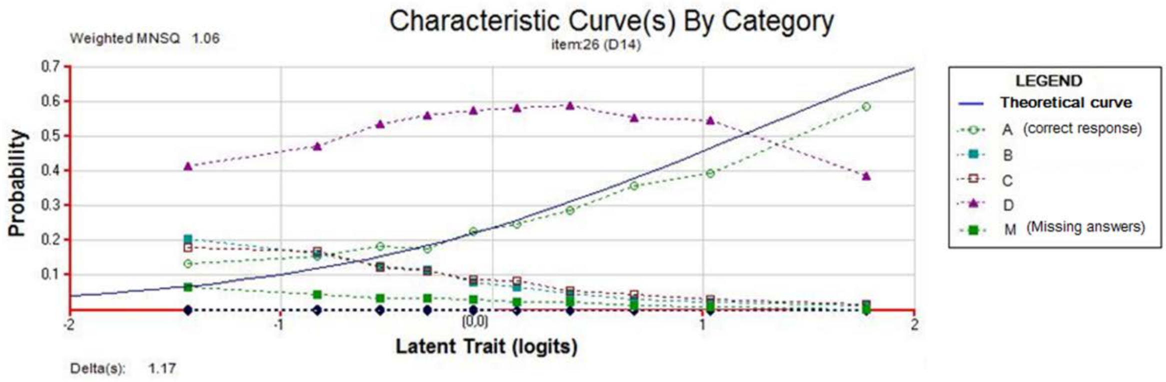


Figure 9. Distractor Plot for item D14 from the grade 6 mathematics test of 2013

One possible explanation for this choice echoes that of the previous item; in fact, it is feasible that students identified in the text the necessary data and operation for replying to the task and did so without checking the situation as modelled in the picture. In this case, the response option comprises the addition of 3cm to each of the photograph measurements.

CONCLUSIONS

In this paper we presented a research based on Italian standardized assessment, namely the INVALSI tests. Our aim is to highlight the potentialities of an item-level analysis of these data, combining both quantitative analysis based on the Rasch Model and qualitative interpretation of the findings through the lenses of math education theories.

Our hypothesis is that standardized testing data do not provide only rankings or scores related to benchmarks; they may provide a huge amount of information about mathematics learnings and feedbacks about teaching/learning processes. This information is contained not only in global scores (referring to the latent trait measured by the statistical models) but also in current phenomena, observed through the single items. According to Leder and Lubienski (2015, p. 35) "Item-level analyses can pinpoint the mathematics that students do and do not know, including which problems most students can and cannot solve, and which problems have the largest disparities between groups. This information can inform both textbook writers and teachers, as they strive to address curricular areas in need of additional attention. Hence, it is important for item-level analyses to be systematically conducted and reported".

In this paper we show that quantitative analysis of some items reveals particular item behaviour; the examples reported show that such behaviours are connected with well-known phenomena in mathematics education research, which are closely linked to classroom practices and the discipline's character. In this perspective, the examples provided highlight how statistical analysis are interesting for mathematics education. In fact, such statistical facts suggest that some well-known phenomena are measurable in terms of students ability in the test. In particular, focusing on *distractor plot* output of the Rasch Models, we identify some items in which the trend of one of the incorrect answers has a particular behaviour (it is more attractive for medium ability levels) and this behaviour can be explained using the didactic contract construct.

In order to confirm the research hypothesis, we carried out a statistical analysis on some INVALSI tasks by tackling different content areas (Numbers, Space and Shape, Relations and Functions) from different scholastic levels (from Primary School to High School).

From a statistical point of view, all the items analysed display good statistical features and are coherent with the Rasch model used for the test analysis (Barbaranelli & Natali, 2005; INVALSI, 2017). Analysing the distractor plots of all the items selected, however, it may be noted that in each there is at least one distractor curve that displays a "humped performance".

Looking for "humped performance" behaviour, we select several item in different mathematical content areas. In the example we present two item in "Numbers" (D14, D5), one in "Space and Shape" (D14) and one in "Relations and Functions" (D7). This fact suggest that such phenomena are not linked with content area but involves other factors linked with teaching and learning practices.

In the same way, we collect item from grade 2 to grade 10; this is totally contrary to the assumption that such factors are typical of low grade students, instead involving students of different school levels. For example, in D5 we show that 15 years-old students are particularly attracted to mnemonically applied calculus tables, just like grade 5 students who need formal justification.

The item were analysed through the lens of mathematics education, and the emerging results point to implicit and explicit rules established in the classroom, especially regarding the didactic contract (Brousseau, 1988) and we observed this particularly for specific students' ability levels.

The parallel between statistical analyses and didactic interpretation of the items allows us to verify the existence of the didactic contract and measure its effects; by analysing the distractor plots it is possible to identify which ability levels are most influenced by these phenomena. In particular, it can be seen that the effects result more evident regarding medium-ability level students. This is completely in line with the nature of the construct used to interpret the phenomena, a notion closely linked to classroom habits and repetition of tasks and resolution methods, in the presence of limited mastery of the content and concepts being used.

This initial study reveals that, regarding the items analysed, the effects of the didactic contract seem to affect particularly students of medium-level ability as opposed to other ability levels: the "humped performance" of the options displaying the phenomena under analysis may be due to the fact that students of low-level ability are not very keen on didactic practices, whilst better students manage to overcome the obstacles facing them thanks to their bond with the didactic method and their teacher. Furthermore, it is important to note that the phenomena encountered are often linked to mistaken knowledge or the result of bad teaching practices; this aspect then is not connected to absence of knowledge or non-participation in classroom activity, common markers of low-level performance in the tests. The close link that these constructs have with classroom practices would appear to confirm the statistical data: further analysis of other types of items covering a wider range of knowledge and mathematical skills could confirm these results.

REFERENCES

Anderson, J. O., Lin, H. S., Treagust, D. F., Ross, S. P., & Yore, L. D. (2007). Using large-scale assessment datasets for research in science and mathematics education: Programme for International Student Assessment (PISA). *International Journal of Science and Mathematics Education*, *5*(4), 591-614. https://doi.org/10.1007/s10763-007-9090-y

Arzarello, F., Garuti, R., & Ricci, R. (2015). The impact of PISA studies on the Italian national assessment system. In *Assessing Mathematical Literacy* (pp. 249-260). Springer, Cham. https://doi.org/10.1007/978-3-319-10121-7_13

Barbaranelli, C., & Natali, E. (2005). *I test psicologici: teorie e modelli psicometrici*. Carocci.

Bolondi, G., Branchetti, L., Ferretti, F., Lemmo, A., Maffia, A., Martignone, F., Matteucci, M., Mignani, S., & Santi, G. (2016). *Un approccio longitudinale per l'analisi delle prove INVALSI di matematica: cosa ci può dire sugli studenti in difficoltà? Report concorso idee per la ricerca*, pp. 81-102. Roma: INVALSI.

Bolondi, G., Cascella, C., & Giberti, C. (2017). *Highlights on gender gap from Italian standardized assessment in mathematics*. Universita Karlova Press.

Branchetti, L., Ferretti, F., Lemmo, A., Maffia, A., Martignone, F., Matteucci, M., & Mignani, S. (2015). A longitudinal analysis of the Italian national standardized mathematics tests. *Proceedings of the 9th Conference of European Research in Mathematics Education*, (pp. 1695-1701) Prague, Czech Republic: Charles University in Prague, Faculty of Education and ERME.

Brousseau, G. (1980). L'échec et le contrat. Recherches, n.41, 177-182.

Brousseau, G. (1988). Le contrat didactique: le milieu. *Recherches en Didactique des Mathématiques, 9*(3), 309-336.

D'Amore, B. (2008). Epistemology, didactics of mathematics and teaching practices. *Mediterranean Journal of Research in Mathematics Education*, *7*(1).

Di Tommaso, M. L., Mendolia, S., & Contini, D. (2016). The Gender Gap in Mathematics Achievement: Evidence from Italian Data. *IZA Discussion paper*, n.10053, Bonn.

EMS-EC (Education Committee of the EMS) (2012). What are the Reciprocal Expectations between Teacher and Students? Solid Findings in Mathematics Education on Didactical Contract. *Newsletter of the European Mathematical Society*, *84*, 53-55.

Ferretti, F. (2015). *L'effetto "età della Terra". Contratto didattico e principi regolativi dell'azione degli studenti in matematica* (Doctoral thesis), Alma Mater Studiorum Università di Bologna. Retrieved from http://amsdottorato.unibo.it/7213/4/Ferretti_Federica_Tesi.pdf https://doi.org/10.6092/unibo/amsdottorato/7213

Ferretti, F., & Gambini, A. (2017). A vertical analysis of difficulties in mathematics by secondary school to level; some evidences stems from standardized assessment. *Proceedings of the 10th Conference of European Research in Mathematics Education*, Dublin (Ireland).

Gestinv 2.0. (07.01.2018). *Archivio interattivo delle prove Invalsi*. Retrieved from http://www.gestinv.it

Giberti, C., Zivelonghi, A., & Bolondi, G. (2016). Gender differences and didactis contract: analysis of two Inalsi tasks on power properties. *40th PME proceedings*, 275.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.

INVALSI. (2017). *Rilevazione nazionale degli apprendimenti 2016-2017. Rapporto tecnico*. Retrieved on March 2018 from http://www.invalsi.it/invalsi/doc_eventi/2017/Rapporto_tecnico_SNV_2017.pdf

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come, *Educational Researcher, 33*(7), 14-26. https://doi.org/10.3102/0013189X033007014

Leder, G., & Lubienski, S. (2015). Large-Scale Test Data: Making the Invisible Visible. In *Diversity in Mathematics Education* (pp. 17-40). Springer International Publishing. https://doi.org/10.1007/978-3-319-05978-5_2

Looney, J. W. (2011). *Integrating formative and summative assessment: progress toward a seamless system?* OECD Education Working Paper 58. OECD Publishing. https://doi.org/10.1787/5kghx3kbl734-en

Mellone, M., Romano, P., Tortora, R., Statale, L. S., Mangino, M. B., & Pagani, S. I. (2013). Different ways of grasping structure in arithmetical tasks, as steps toward algebra. In *Proceedings of CERME* (Vol. 8, pp. 480-489).

Middleton, J. A., Cai, J., & Hwang, S. (2015). Why mathematics education needs large-scale research. In *Large-scale studies in mathematics education* (pp. 1-13). Springer International Publishing. https://doi.org/10.1007/978-3-319-07716-1_1

Rasch G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*, Danmarks Paedagogiske Institut, Copenhagen.

Sfard, A. (2005). What could be More Practical than Good Research? *Educational Studies in Mathematics*, *58*(3), 393-413. https://doi.org/10.1007/s10649-005-4818-5

http://www.ejmste.com