

A mixed approach to interpret large-scale assessment psychometric results of the learning of mathematics

Un approccio misto all'interpretazione dei risultati psicometrici delle valutazioni su larga scala dell'apprendimento della matematica

Giorgio Bolondi¹ e Clelia Cascella²

¹Faculty of Education, Free University of Bozen-Bolzano, Italy

²INVALSI, Italy

Abstract. *In this paper we present a case for arguing that psychometric properties of a mathematical item administered in a large-scale assessment (even “bad” psychometric properties) may be usefully connected to the didactic features of the task proposed and to the behaviour of the students. We analyse 354 items administered to grade 10 students in the frame of the Italian national evaluation system. This approach may provide interesting issues for educational research.*

Keywords: large-scale assessment, Rasch model, task design.

Sunto. *In questo articolo presentiamo un caso per mostrare come le proprietà psicometriche di una domanda di matematica somministrata in una valutazione su larga scala (anche proprietà “brutte”) possano essere utilmente messe in relazione con le caratteristiche didattiche della consegna e con il comportamento degli studenti. A questo scopo analizziamo 354 domande somministrate agli studenti di grado 10 nel contesto delle prove nazionali di valutazione in Italia. Questo approccio può fornire interessanti prospettive per la ricerca didattica.*

Parole chiave: valutazioni su larga scala, modello di Rasch, task design.

Resumen. *En este artículo presentamos un caso donde se muestra cómo las propiedades psicométricas de una pregunta de matemática dentro de un proceso de evaluación a gran escala (incluso las propiedades “negativas”) pueden ser relacionadas positivamente con las características formativas de la pregunta y con el comportamiento de los estudiantes. Para esto, analizamos 354 preguntas propuestas a estudiantes de grado 10 en el contexto de las pruebas nacionales de evaluación en Italia. Este enfoque puede proporcionar perspectivas interesantes para la investigación didáctica.*

Palabras claves: evaluación a gran escala, modelo de Rasch, diseño de tareas.

1. Introduction

The debate on if and how a psychometric approach may provide useful information for educational research, and for mathematics education research in particular, is long-standing and yet animated (Primi, 2017). A general discussion on the use of large-scale assessment (LSA) data in research in mathematics education is reported in Meinck, Neuschmidt, and Taneva (2017), and De Lange (2007), where the information content of quantitative data is analysed. A major problem in the application of psychometric results related to large-scale assessment to the actual study of educational problems is the qualitative description and interpretation of the latent trait, that is to say of “what” is really measured by the assessment. One must consider this “what” in terms of its relationships with the epistemological status of the discipline involved and of the intended school curriculum.

The first concern of psychometrists is the fitting of the empirical data with the statistical model chosen. This is fundamental, but of course researchers in education are more interested in how far the model provides information about the reality. Our research hypothesis is that a moderate item misfit do not need be interpreted necessarily as a limitation (of the test or even of the choice of the model), but as a potential source of information. We support this approach with the study of a case coming from the Italian national large-scale assessment in mathematics. We consider a group of nearly-misfitting items with similar psychometric features which can be linked to possible behaviours of students, due to classroom habits. Hence this moderate misfit throws lights over the relationship between the latent trait and the intended construct to be assessed (let us call it “mathematics learning according to the Italian curricula”).

Although a lot of different methods and techniques have been proposed within the item response theory (IRT) to assess item fit, this topic still raises relevant questions to which not completely satisfactory answers have been given (e.g. Hattie, 1984, 1985; Embretson & Reise, 2000), especially regarding large-scale data (Gustafson, 1980). In such cases, in fact, the real objective to pursue nowadays is a definition of the extent to which deviation between expected and observed values can be considered tolerable. This objective cannot be completely achieved in the absence of specific hypotheses addressing the issue in order to understand (at least in part) the causes of the observed violation. The general issue that we tackle here is the proposal of a solution to this lack, via a mixed method based on the interpretation of misfitting item behaviour from a didactical point of view. In general, a misfitting behaviour is seen as problematic for the application and issues of analysis. The background assumption of this paper is that a misfitting behaviour of one or more items may contain information both about the actual construct measured by the test, and the misfitting item itself.

In this paper, as an introductory example of this approach, we present a

close examination of one violation type, over-discrimination, which allows us to highlight some characteristics of items that, on the one hand, clarify the functionality of the employed statistical model and, on the other, offer some interesting avenues to explore regarding the nature of the phenomenon and spread some lights on students' behaviour.

In our investigation, besides the standard parameters used within the statistical model, fit control is based on the graphical inspection of item characteristic curves (ICCs), estimated by using the Rasch model (one of the most popular tools in educational research to estimate students' ability), which allows the identification of deviations between observed and expected values for specific ability levels. This can be particularly useful in order to formulate specific hypotheses aimed at understanding and identifying some possible causes of violations.

Our methodology may be framed as a mixed paradigm: we perform a qualitative interpretation of large quantitative data, analysed by means of the Rasch model and collected by the Italian national institute for the evaluation of the education system (INVALSI) to assess students' ability both in maths and text comprehension, yearly, at grade 2 and 5 (primary school), at grade 8 (lower secondary school) and at grade 10 (upper secondary school).

We individuate a group of nearly-misfitting items with similar psychometric properties (in particular, a remarkable over-discrimination) whose specific common features suggest possible causality sequences and point to possible behaviours of students, due to classroom habits. We point out a possible relationship between psychometric properties of the results (the over discrimination), features of the item (distance from "standard" classroom tasks), and behaviour of students facing the task (in particular, of "low-achieving" students). Of course, these conjectures need to be validated through ad hoc experiments and qualitative methods, and this is what we consider really interesting of this approach, in the light of the discussion mentioned above: the analysis of quantitative evidences, driven by suitable theoretical lenses, can suggest valuable research paths.

In the discussion and in the conclusion sections, we explain how these results promote a general discussion of the interpretation of misfit. Finally, we underline that our case study is based on Italian data and we make explicit reference to features of Italian didactics praxis; nevertheless, our results show that our approach may provide a new light on the general interpretation of standard Rasch outputs, which are commonly used in educational research all over the world.

2. Item fit control within the Rasch analysis

In educational research, one of the most extensively employed models for the estimation of abilities and competences is the Rasch model (Rasch,

1960/1980). It belongs to the item response theory (IRT) and hypothesizes that a subject's answer to an item depends on his/her relative ability, i.e. the subject's intrinsic ability as compared with the difficulty of the item. Therefore, it is able to scale both subjects and items along the same latent trait (in our case, math ability) depending on students' ability and items' difficulty, respectively.

The popularity of the Rasch model is due both to its intrinsic ease of use and its statistical properties (such as measurement invariance and specific objectivity), strongly desirable especially in large-scale assessment. These properties hold only if both the item and test functionality are consistent with theoretical assumptions underlying the Rasch model (Hambleton & Swaminathan, 1985):

- 1) one-dimensionality (all items refer to a unique – or at least prevalent – latent dimension);
- 2) local independence (the probability of correctly answering an item is stochastically independent from the probability of correctly answering any other item within the same test);
- 3) monotonicity (the probability of a correct answer increases proportionally to student ability).

Moreover, although conceptually distinct, one-dimensionality and local independence are strictly interrelated because, by definition, data are unidimensional when the answers are locally independent in relation to the same latent trait (McDonald, 1981).

Although this topic is extremely relevant because the interpretability of results depends on its properties, this issue is still topical because it nowadays raises relevant questions wider than the method itself. In fact, over the years, a lot of different methods and techniques have been proposed to judge how well an IRT model represents data at the item, person, or model levels (for an updated overview, see for example Wu and Adams, 2013), but there are no procedures that result in research stating definitively that a particular model does or does not fit, or is or is not appropriate (Hambleton & Swaminathan, 1985).

For these reasons, IRT model-fit assessment is still an active area of research to this day, in particular for big data, where adequate fit level can rarely be observed (Wright, Linacre, Gustafson, & Martin-Löf, 1994). In fact,

Though the ideal for measurement construction is that data fit the Rasch model, all empirical data departs from the model to some extent. (...) Conventional statisticians base their decisions on significance tests, but these are heavily influenced by sample size. (Wright et al., 1994, p. 370)

In fact, “no model can ever be supposed to be perfectly fitted by data, so with a sufficiently large sample any model would have to be discarded” (Gustafson, 1980, as cited in Wright et al., 1994, p. 370). On the contrary,

for large sets of data it is too destructive to let an ordinary significance test decide whether or not to accept a proposed statistical model [or data], because, with few exceptions, we know that we shall have to reject it even without looking at the data simply because the number of observations is so large. (Martin-Löf, 1974, p. 3)

In other words, especially for large data set, the real question is: How much unmodelled noise is tolerable?

In light of this, the aim of this paper is not to propose another method/technique to assess item (mis)fit but rather our objective here is to propose an alternative use, for researcher, practitioners and teachers as well, of traditional Rasch output as the basis of theoretical reflections about causes of item misfit. In this sense, we refer to the practical assessment of item fit in real situations, in relation to which a scientist often must use “his or her best judgment” (Embretson & Reise, 2000, p. 233).

For these reasons, our general question is: How can Rasch analysis help educational researcher in developing his/her judgment?

3. Methods and materials

3.1. Methodology

It is absolutely unquestionable that different methods have different strengths and weaknesses, and that a qual-quant approach often offers a way to bring different strengths together in the same research (Morgan, 1998) and to overcome some weaknesses of one technique by the help of another. In particular, in the present study, we combine quantitative evidence derived from the statistical Rasch model with a qualitative interpretation (by the means of didactical lens) in a case study, i.e. a methodological approach based on the combination of different techniques with “the purpose of illuminating a case from different angles (...) focusing on a particular phenomenon might be *read* as an investigation of a different phenomenon” (Johansson, 2003, pp. 3–6).

In fact, although the Rasch model is a very useful tool for the analysis of students’ ability, we think that the interpretation of both persons’ and items’ parameters along the same latent trait should be broader than that commonly applied. We refer in particular to the interpretation of fit analysis. As pointed out in the previous paragraph, different fit measures exist within Rasch analysis and they are particularly relevant in order to understand if and how data fit the Rasch theoretical assumptions. From our perspective, deviations from a model’s expectation (i.e. misfit) may provide significant information about students’ answering strategy behaviour and, therefore, they can support our research to formulate some specific hypotheses about mechanisms underlying items’ misfit.

3.2. Data collection

In this study, we use data collected by the Italian national institute for the evaluation of the education system (INVALSI), from 2010 up to 2017, at grade 10 (second year of upper secondary school, 15- or 16-year-old students). Each year, the INVALSI system administers a math achievement test aimed at assessing mathematical competence, i.e.

the ability to develop and apply mathematical thinking in order to solve a range of problems in everyday situations. Building on a sound mastery of numeracy, the emphasis is on process and activity, as well as knowledge. Mathematical competence involves, to different degrees, the ability and willingness to use mathematical modes of thought (logical and spatial thinking) and presentation (formulas, models, constructs, graphs, charts). (European Parliament and Council, 2006, p. 15)

In line with this definition, all math achievement tests administered by INVALSI are aimed at exploring this specific ability as clearly outlined and more extensively detailed by the *Quadro di riferimento* (framework of reference) and *Indicazioni Nazionali per il curriculum* (National Guidelines for the Curriculum), both of which are available on the INVALSI website.

From 2010 up to 2017, INVALSI has administered 354 math test items at grade 10, at census level and, at the same time, to randomly selected classes on a national scale. In these classes, test administration is supervised by an external examiner delegated by INVALSI in order to guarantee the fairness of administration procedure and to prevent cheating. Thus, for the purposes of the present study, we analyse sample data. Case numerosity is shown in Table 1.

All these items have been analysed and classified depending on their psychometric functionality. In fact, as previously pointed out, for big data at least, “no model can ever be supposed to be perfectly fitted by data” (Gustafson, 1980, as cited in Wright et al., 1994, p. 370). Therefore, in order to assess item fit, we need to refer to tolerance intervals within which deviation between observations and expectations is not unproductive for measurement. Simulation studies on this topic have produced a general tolerance interval (i.e. productive for measurement), ranged into the interval [0.80; 1.20] (Wright et al., 1994; Linacre, 2002; INVALSI, 2015, 2016, 2017). Nevertheless, according to the authors, critical tolerance intervals should not be defined a priori. Thus, we have used big INVALSI data to adapt Wright’s intervals to the Italian data. To this end, we have analysed and classified all data collected by INVALSI, year by year, from 2010 up to 2017 (354 items) according to their psychometric functionality (normal/misfitting).

Table 1
Students involved year by year

		Case numerosity	N. of test items
2010–2011	Population	385411	53
	Sample	11.7%	
2011–2012	Population	413847	54
	Sample	10.1%	
2012–2013	Population	418243	50
	Sample	9.1%	
2013–2014	Population	410609	54
	Sample	9%	
2014–2015	Population	288248	53
	Sample	9.5%	
2015–2016	Population	349263	52
	Sample	8.9%	
2016–2017	Population	392396	59
	Sample	8.8%	

Source: our elaboration on INVALSI data

We have observed three main intervals (type A, B, and C), as shown in Table 2, according to items' weighted MNSQ, a typical fit measure within the Rasch framework. In INVALSI dataset, we have observed three tolerance empirical intervals and have classified them in three categories. The first and the second intervals, respectively equal to [0.95; 1.05] (type A) and [0.90; 1.10] (type B), comprise around 70% of items administered in each achievement test and both guarantee fair parameter estimations. The third interval (type C) contains items with weighted fit index lower than 0.90 or higher than 1.10.

Table 2
Misfitting items by category

Item type	Number of items (from 2010 to 2017)
Type A (INFIT ranged [0.95; 1.05])	148 (42% over 354 items)
Type B (INFIT ranged [0.90; 0.95[U]1.05; 1.10])	97 (27% over 354)
Type C (INFIT lower than 0.90 or higher than 1.1)	111 (31% over 354)

Source: Our elaboration on INVALSI data

In the present study, we focus our attention on over-discriminating items. In particular, we have selected them according to two criteria: 1) weighted MNSQ lower than 0.90; and, 2) percentage of missing values equal to or greater than 20%. This percentage has been observed especially at the bottom of the latent trait (i.e. in correspondence with lower-ability levels) but we have observed the combination of both features irrespectively of an item's difficulty. Thus, items selected here are characterized by high, medium, and low difficulty, in the same percentage. From 2010 to 2017, from among over 354 items administered by INVALSI, 35 matched the first criterion (weighted MNSQ lower than 0.90), 15 of which also matched the second criterion (percentage of missing values equal to or greater than 20%). We name this group of items o-DM (in order to collect all main features of selected items, i.e. their over-discrimination and the high percentage of missing values). In order to propose an in-depth investigation of our selected items, we select and present items in the following paragraph 4 (Feagin, Orum, & Sjöberg, 1991). Over-discrimination can be caused by different factors, such as response dependence (i.e. a violation of item independence as requested by the Rasch model, according to which a student's answer depends on his/her response to one or more previous items) but it can also be caused by other factors, such as item formulation. All o-DM items do not violate the independence assumption and do not seem to be affected by the factors cited above.

3.3. *The graphical inspection of the item characteristic curve*

In order to interpret all discrepancies between observed and expected item behaviour, we present a fit analysis based on graphical inspection of the item characteristic curves (ICCs).

The graphical inspection of ICC allows the identification of different answering behaviour profiles. In fact, each ICC is a logistic regression line, with item performance regressed on examinee ability, which links a student's probability of success on an item to the trait measured by the set of test items. The probability of a correct answer is estimated by comparing the student's intrinsic ability (i.e. the level of skill held by the student) and item difficulty (i.e. the level of skill that the student must have to correctly answer). As previously said, since the Rasch model hypothesizes that just student ability

and item difficulty determine the interaction between person and item, it generates a very sturdy estimation environment “against which to test data for the presence of anomalous behaviour that may influence the estimation of item and person parameters. This identification (...) addresses any potential measurement disturbance” (Smith, 1993, p. 262). For this reason, the graphical inspection of the ICC seems a very adequate method for the study of discrepancy between estimated and observed probability of correctly answering an item. In this sense, the observation of such deviations allows control of the consistency between the dataset and model’s theoretical assumptions, with particular attention to cumulativity, according to which students with higher ability levels must have a higher probability of correctly answering a higher number of items than other students.

Since different software produce different outputs, we have used and compared both RUMM2030 and ConQuest 4.0. However, in this study, they show very similar results. Thus, in the following paragraph, we present output provided by ConQuest 4.0. Since ConQuest provides output belonging to both the classical test theory and the Rasch analysis, for each item, we report:

- *Discrimination index.* In addition to Rasch analysis, ConQuest performs also a traditional item analysis for all of the generalised items, including item discrimination index. In particular, ConQuest provides the item point-biserial (discrimination) as a correlation between the item scores and the test scores, formed by all items in the test or formed by all of the rest items in the test (Le, 2012). In the light of the classical test theory, items showing a high point-biserial coefficient differentiate subjects depending on their ability better than items showing a lower coefficient (Barbaranelli, Natali, 2005). In other words, the point-biserial coefficient measures the relationship between the item score and the test score: the higher the correlation is, the higher the item validity is (i.e., its capacity to reveal the latent trait). Therefore, we interpret the point-biserial correlation by category as an index of consistency between students’ answering behaviour (observed in each answer group) and the overall test.
- *Weighted MNSQ.* A synthetic index belonging to the Rasch framework, used to assess fit between data and model theoretical assumptions. The weighted MNSQ, developed by Rasch (1960/1980) and Wright (1977), can test violations of the model assumptions of no-guessing and homogeneity of item discrimination. It has expectation 1.0, and range from 0 to infinity. Mean-squares greater than 1.0 indicate underfit to the Rasch model, i.e., the data are less predictable than the model expects. For example,

a mean-square of 0.7 indicates a 30% deficiency in Rasch-model-predicted randomness (i.e., the data are too Guttman-like), which implies $100 \cdot (1 - 0.7) / 0.7 = 43\%$ more ambiguity in the inferred measure than modelled (e.g., the item difficulty estimated from low-ability persons differs noticeably from the item difficulty estimated from high-ability persons). (Wright et al., 1994,

p. 370)

Mean-squares less than 1.0 indicate overfit to the Rasch model, i.e., the data are more predictable than the model expects. For example, “A mean-square of 1.2 indicates that there is 20% more randomness (i.e., noise) in the data than modelled” (Wright, et al. 1994, p. 370). Items showing a weighted MNSQ minor than 1.00 are called over-discriminating and we focus our attention on them.

- *Item characteristic curve* (ICC). Coherently with Rasch theory, ICC expresses the estimated probability of successfully endorsing an item while taking into account students’ relative ability and the distribution of observed correct answers provided by students clustered by their own ability level. The ConQuest graphical output plots, into the same graph, 1) the observed probability of correctly answering an item by category (i.e., A – B – C – D for multiple-choice items or 0 – 1 for binary items); and, 2) missing values.

4. Results

In this paragraph, we show our item selection, in particular those over-discrimination with a high percentage of missing values, in particular at the lower end of the latent trait, selected over the 354 administered between 2010 and 2017.

A paradigmatic example of our analysis is question D8 administered in 2011 to grade 10 students (hence, attending the 2nd grade level of upper secondary school- 15- or 16-year-old students) (Figure 1).

D8. La dimensione di un televisore è la misura della diagonale dello schermo espressa in pollici (1 pollice = 2,54 cm). Nei televisori di nuova generazione il rapporto tra la larghezza e l'altezza dello schermo è 16:9.

Se la larghezza dello schermo di uno di questi televisori è circa 57,5 cm, qual è all'incirca la sua altezza?

Risposta: cm

Authors' translation

TV size is the measure of its diagonal in inches (1 inch = 2.54 cm). For new generation televisions, the relationship between height and width is 16:9.

- a. If TV width is 57.5 cm, what is its height?

Answer: _____

Figure 1. Item D8 administered by INVALSI in 2011 at grade 10.

Item D8 is an open-answer item as shown in the following figure. Its content is related to proportionality. The task is proposed in a real context and students were allowed to use a calculator. In Figure 2, we report the item characteristic

curve, by category (i.e. a line for each answer option and for missing values – coded 7 and 9 – is drawn into the same graph). In the figure, the blue line represents the model probability of giving the correct answer (i.e. the probability of correctly answering this item as estimated by the Rasch model for all students for each ability level) (Figure 2).

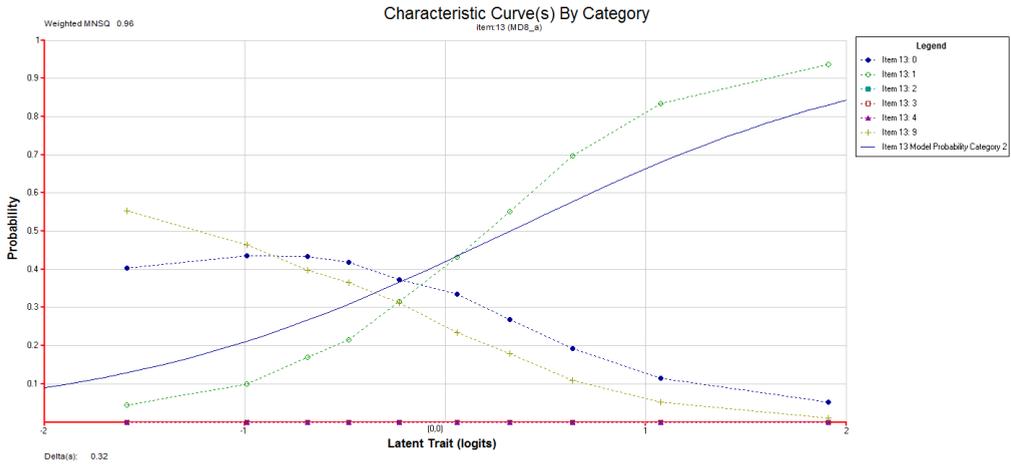


Figure 2. Item characteristic curve by category, item D8 administered by INVALSI in 2011 at grade 10. Note. D8 is open-ended item. Answer categories for this item are: 0 = wrong answer; 1 = correct answer; 9 = missing. Therefore, categories 2, 3, and 4 (included by ConQuest by default) do not represent any answer option and thus these categories are represented by the lines overlapping the x-axis.

In Table 3, we report some descriptive information provided by ConQuest. For item D8, we process 43,458 cases. This item shows high discrimination, according to the classical test theory (i.e. the item’s power to differentiate students according to their math competence), equal to 0.63 (Alagumalai & Curtis, 2005). Its item parameter estimated by the Rasch model is equal to +0.33 (i.e., medium difficulty level). For all items, in fact, the estimated student ability/item difficulty are scaled on the same latent trait, empirically ranged in the interval $[-2; +2]$ logit, where 0.00 represents medium ability/difficulty level; negative parameters (progressively) indicate low ability/difficulty levels; and, positive estimates (progressively) indicate high ability/difficulty levels. For this item, the weighted MNSQ index is less than 1.00 (0.82). It highlights an over-discrimination, but without negative consequences on parameter estimations according to tolerance intervals proposed by Wright et al. (1994) and Linacre (2002), and subsequently employed by the INVALSI to assess item fit (for further details, see INVALSI, 2015, 2016, 2017). Finally, a case count (and relative percentages)

reported for each answer option show a high percentage of missing values (26.04%), especially in the low end of distribution as shown in the previous figure, and a high point-biserial correlation, by answer category. The latter is a common index used in item traditional statistics to assess item quality and it refers to the degree to which an item differentiates correctly (i.e. neither more nor less than what is predicted by the model) among examinees in the behaviour that the test is designed to measure. Typically, for both multiple choice and open-answer items, we observe a positive point-biserial discrimination for the correct answer and negative point-biserial discrimination for each of its distractors or wrong answers (Millman & Green, 1989). In this case, we observe a high point-biserial coefficient for both correct answer (Pt Bis = 0.63) and missing values (Pt Bis = -0.40) (Table 3).

Table 3

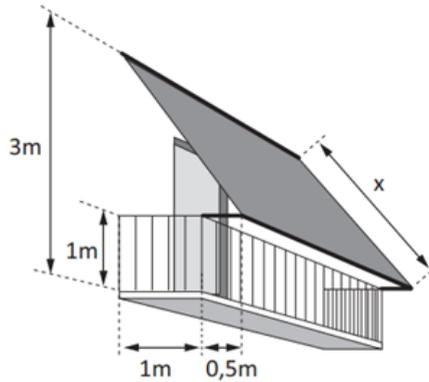
Item analysis, Item D8 administered by INVALSI in 2011 at grade 10

Item	D8							
Cases for this item	43458							
Weighted MNSQ	0.82							
Item parameter	0.33							
Label	Score	Count	% of total	Pt. Bis	t (p)	PV1Ag:1	PV1 SD:1	
0	0.00	13076	30.09	-0.28	-61.95(.000)	-0.40	0.75	
1	1.00	18606	42.81	0.63	168.34(.000)	0.67	0.87	
7	0.00	197	0.45	-0.04	-7.72(.000)	-0.55	0.76	
9	0.00	11579	26.64	-0.40	-91.53(.000)	-0.62	0.74	

Note. 0 = wrong answer; 1 = correct answer; 7 = not valid answer (treated as missing); 9 = missing answer.

A similar item functionality can be observed for question D24, as administered by INVALSI in 2012 (Figure 3).

D24. Occorre confezionare una tenda da sole per il balcone in figura.
 La tenda deve essere fissata al muro a 3 m di altezza dal pavimento del balcone che è largo 1 m.
 La tenda deve sporgere 0,5 m dalla ringhiera che è alto 1 m.



Scrivi i calcoli che fai per trovare la lunghezza x della tenda e infine riporta il risultato.

.....

Risultato: $x = \dots\dots$ metri

Authors' translation:
 We want to make an awning for the balcony drawn in the figure.
 The awning has to be fixed on the wall at 3 meters off the balcony floor that is 1 meter wide. The awning has to stick 0.5 meter out the balustrade (that 1 meter high).

Figure

What is the awning length (x)? Write the calculation and then the result.

.....

Result: $x = \dots\dots$ meters

Figure 3. Item D24 administered by INVALSI in 2012 at grade 10.

Item D24 is a multiple-choice item. The item's intent here is to explain the different steps leading to application of the Pythagorean theorem in a real context (Figure 4).

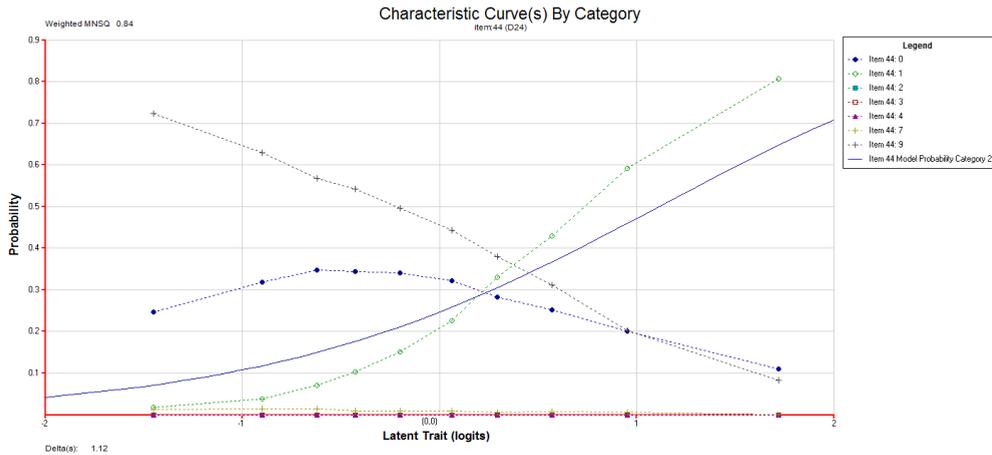


Figure 4. Item characteristic curve by category, item D24 administered by INVALSI in 2012 at grade 10. *Note.* D24 is open-ended item. Answer categories for this item are: 0 = wrong answer; 1 = correct answer; 9 = missing. Therefore, categories 2, 3, and 4 (included by ConQuest by default) do not represent any answer option and thus these categories are represented by the lines overlapping the x-axis.

For item D24, we process 41,812 cases. This item has a high discrimination index, equal to 0.59. Its parameter (item difficulty) is high on the Rasch scale (equal to +1.15). The INFIT measure is less than 1.00 (equal to 0.84) and indicates a strong item over-discrimination, without causing any significant bias in parameters estimation (according to Linacre, 2002; Wright et al., 1994). Case count (and relative percentages) reported for each answer option show a very high percentage of missing values (43.79%), especially in the low end of distribution as shown in the previous figure. Finally, point-biserial coefficient by answer category shows a strong correlation between students' answering behaviour for both the correct answer (Pt Bis = +0.59) and the missing values (Pt Bis = -0.41). This means that students' answering behaviour in these categories is consistent to what the test is designed to measure (math competence). Instead, point-biserial correlation calculated for the wrong category is very low (Pt Bis = -0.13), thus suggesting that something different from the measured construct might lead students to provide a wrong answer (Table 4).

Table 4

Item analysis, Item D24 administered by INVALSI in 2012 at grade 10

Item	D24						
Cases for this item	41812						
Weighted MNSQ	0.84						
Item parameter	1.15						
Label	Score	Count	% of total	Pt. Bis	t (p)	PV1Ag:1	PV1 SD:1
0	0.00	11555	27.64	-0.13	-26.80(.000)	-0.19	0.83
1	1.00	11562	27.65	0.59	148.86(.000)	0.89	0.85
7	0.00	386	0.92	-0.04	-8.79(.000)	-0.42	0.94
9	0.00	18309	43.79	-0.41	-90.62(.000)	-0.44	0.85

Note. 0 = wrong answer; 1 = correct answer; 7 = not valid answer (treated as missing); 9 = missing answer.

Another interesting case is item D17 administered by INVALSI in 2014 (Figure 5).

D17. Data l'equazione $(2k-3)x+1-k=0$, in cui x è l'incognita e k è un numero reale.
La soluzione dell'equazione è 1 per k =

Authors' translation:
It is given the following equation $(2k-3)x+1-k=0$, where x is the unknown variable and k is a real number.
The solution of the equation equals 1 for k =

Figure 5. Item D17 administered by INVALSI in 2014 at grade 10.

Item D17 is an open-answer item, and differently from the previous ones it is a non-contextualized item, but it is not a standard task in Italian school: given the solution, the student has to determine the parameters of the equation, whilst, usually, he/she has to determine the solution (Figure 6).

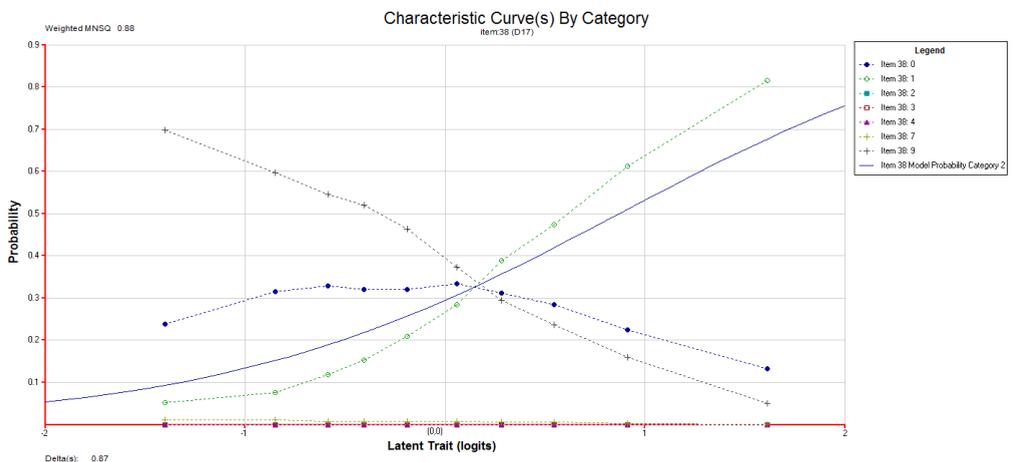


Figure 6. Item characteristic curve by category, item D17 administered by INVALSI in 2014 at grade 10. Note. D17 is open-ended item. Answer categories for this item are: 0 = wrong answer; 1 = correct answer; 9 = missing. Therefore, categories 2, 3, and 4 (included by ConQuest by default) do not represent any answer option and thus these categories are represented by the lines overlapping the x-axis.

For item D17, we processed 36,932 cases. Similarly, to previous items, item D17 has a high discrimination index, equal to 0.57, and a medium-high level of difficulty on the Rasch scale (equal to +0.87). The weighted MNSQ is low (0.88) and indicates a strong item over-discrimination, without causing significant bias in parameters' estimation (Linacre, 2002; Wright et al., 1994). Finally, the case count (and relative percentage) reported for each answer option shows an insignificant percentage of inadmissible answers (less than 1%), a substantial equal distribution between wrong and correct answer, and very high percentage of missing values (39.38%), especially in the low tail of distribution as shown in the previous figure. Finally, point-biserial coefficient by answer category shows a strong correlation between students' answering behaviour for both the correct answer (Pt Bis = +0.50) and the missing values (Pt Bis = -0.41). This means that students answering behaviour in these categories is consistent to what the test is designed to measure (math competence). Instead, point biserial coefficient calculated for the wrong category is very weak (Pt Bis = -0.07), probably suggesting that something different from the measured construct might lead students to provide a wrong answer (Table 5).

Table 5

Item analysis, Item D17 administered by INVALSI in 2014 at grade 10

Item	D17						
Cases for this item	36932						
Weighted MNSQ	0.88						
Item parameter	0.87						
Label	Score	Count	% of total	Pt. Bis	t (p)	PV1Ag:1	PV1 SD:1
0	0.00	10379	28.1	-0.07	-12.65(.000)	-0.110	0.720
1	1.00	11748	31.81	0.5	110.74(.000)	0.630	0.790
7	0.00	261	0.71	-0.03	-6.35(.000)	-0.330	0.730
9	0.00	14544	39.38	-0.41	-86.35(.000)	-0.430	0.690

Note. 0 = wrong answer; 1 = correct answer; 7 = not valid answer (treated as missing); 9 = missing answer.

The following item D20 (Figure 7), administered by INVALSI in 2014, requires the solution of a task in a real context. It demonstrates very similar behaviour to the previous item, with a high percentage of missing values, although the item parameter (i.e. its estimated difficulty) is not very high (0.61) (Figure 8).

D20. Da un controllo di qualità è emerso che una macchina ha prodotto 14 pezzi difettosi su una produzione di 1200 pezzi. Che stima è ragionevole fare del numero di pezzi difettosi su una produzione di 2150 pezzi?

Scrivi i calcoli che hai fatto per trovare la risposta e poi riporta il risultato.

.....

Risultato:

Authors' translation:

A quality control revealed that a machinery has produced 14 defective products out of 1,200 products produced. What is the reasonable estimate of defective products when we produce 2,150 products?

Write the calculations you need to answer the question and then write the result.

.....

Results:

Figure 7. Item D20 administered by INVALSI in 2014 at grade 10.

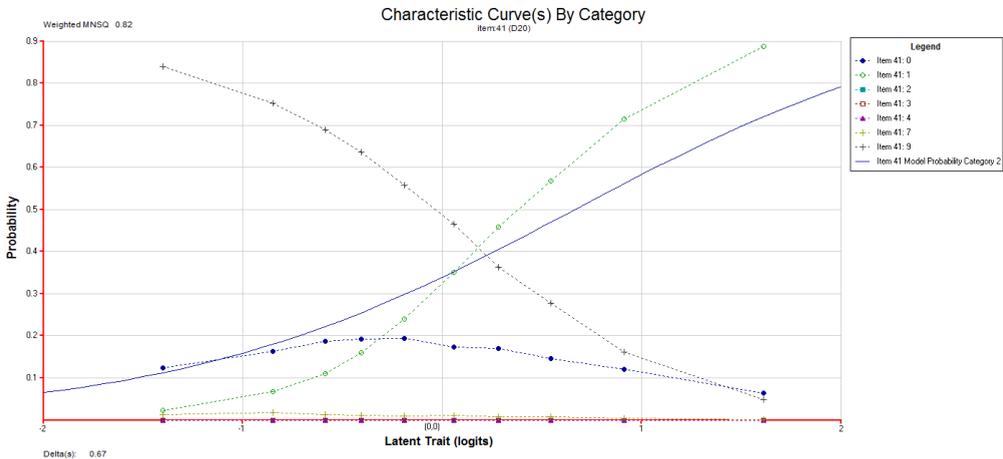


Figure 8. Item characteristic curve by category, item D20 administered by INVALSI in 2014 at grade 10. Note. D20 is open-ended item. Answer categories for this item are: 0 = wrong answer; 1 = correct answer; 9 = missing. Therefore, categories 2, 3, and 4 (included by ConQuest by default) do not represent any answer option and thus these categories are represented by the lines overlapping the x-axis.

For item D20, we process 36,932 cases. This item has a high discrimination index, equal to 0.61. The weighted MNSQ is low (0.82) and thus indicates a strong item over-discrimination, without causing significant bias in parameters estimation (according to Linacre, 2002; Wright et al., 1994). Also in this case, we observe a very high percentage of missing values (47.94%), especially in the low end of distribution as shown in the previous figure. This is not one of the most difficult items within the achievement test and this percentage does

not seem attributable to other factors. In a similar manner to the previous items, point-biserial coefficient by answer category shows a strong correlation between students' answering behaviour for both the correct answer (Pt Bis = +0.57) and the missing values (Pt Bis = -0.51).

This means that students' answering behaviour in these categories is consistent to what the test is designed to measure (math competence). Instead, point-biserial correlation calculated for the wrong category is very weak (Pt Bis = -0.05), probably suggesting that something different from the measured construct might lead students to provide a wrong answer (Table 6).

Table 6

Item analysis, Item D20 administered by INVALSI in 2014 at grade 10

Item	D20						
Cases for this item	36932						
Weighted MNSQ	0.82						
Item parameter	0.67						
Label	Score	Count	% of total	Pt. Bis	t (p)	PV1Ag:1	PV1 SD:1
0	0.00	5666	15.34	-0.05	-9.95(.000)	-0.120	0.700
1	1.00	13212	35.77	0.57	134.92(.000)	0.650	0.740
7	0.00	348	0.94	-0.04	-8.47(.000)	-0.400	0.660
9	0.00	17706	47.94	-0.51	0.00(.000)	-0.440	

Note. 0 = wrong answer; 1 = correct answer; 7 = not valid answer (treated as missing); 9 = missing answer.

5. Discussion

I quesiti sono stati somministrati a 343 alunni sui 349 inizialmente coinvolti; alcuni non li hanno potuti risolvere per vari motivi contingenti.

By employing our methodological strategy to all answers to the items administered at grade 10 by INVALSI year by year, from 2011 to 2017, we individuated a group of items whose weighted MNSQ can be deemed acceptable (i.e. between 0.80 and 0.90) although not completely coherent to Rasch theoretical assumptions. We performed a qualitative analysis of these items, in particular of those showing the most relevant over-discrimination between the 354 items of the dataset.

None of these items present problems regarding the formulation (task posing), nor do they violate the local independence assumption. These items tackle different mathematical contents, and they feature a wide range of difficulty – from medium to high (with item parameter from 0.33 to 1.44, on the Rasch scale, in their tests).

Nevertheless, they show common features in the formulation and type (the input): they are open-ended items, and they can require either a single short answer (such as the result of an arithmetic operation) or a longer text passage to discuss and propose a solution. They are items in a context (real or

mathematical), where the correct answer does not follow from a direct application of a knowledge or a procedure. What is required, instead, is the skill of recognizing in an unusual situation (for Italian praxis) the underlying well-known mathematical concepts. They are related to the process that the OECD-Pisa framework calls “formulating”:

The word “formulate” in the mathematical literacy definition refers to individuals being able to recognise and identify opportunities to use mathematics and then provide mathematical structure to a problem presented in some contextualised form. In the process of formulating situations mathematically, individuals determine where they can extract the essential mathematics to analyse, set up, and solve the problem. They translate from a real-world setting to the domain of mathematics and provide the real-world problem with mathematical structure, representations, and specificity. They reason about and make sense of constraints and assumptions in the problem. (OECD, 2013, p. 28)

When we compare this task with the didactic praxis of Italian schools, we can see that this kind of task is not common (D’Amore, 2014).

These items also have common features in the output, as synthesized by the distractor plot given by ConQuest. They have a weighted MNSQ between 0.80 and 0.90: the model predicts more correct answers for low-ability students and fewer correct answers for high-ability students than those actually observed. Moreover, they are characterized by a very high number of missing answers, especially for students with low ability, and consequently a very low guessing level – far lower than that predicted by the model. We may conjecture a causal relationship between these facts: the high number of missing answers for low-ability students seem to cause a zeroing of the guessing. These “missing” correct answers contribute to an overestimation of the difficulty parameter (as estimated by the model on the whole population). High-ability students, therefore, provide more correct answers than predicted.

We may interpret these facts on the assumption that student behaviour may be explained also as dependent on the specific features of the task: its typology, its relationship with classroom habits, but also the processes involved in answering. A natural interpretation of these facts as a whole is that low-ability students, when facing a task where learned facts and acquired abilities are not immediately recognizable, feel displaced and are more inclined to non-answering. For high-ability students this does not happen, and the items are in fact easier than those in the prediction of the model, hence the over-discrimination. We observe then that this over-discrimination – the quantitative datum given back by Rasch analysis – is related via our qualitative analysis to this low-ability students’ displacement, when facing a mathematization situation. Symmetrically, the capability of mathematically formulating a situation seems to be an element of strong discrimination between students.

6. Conclusion and further issues

The study performed in this paper is based on a wide empirical dataset of mathematical items, gathered from the huge heritage of information collected by the INVALSI system over the years.

Among the items administered, there are some which are correct in formulation, coherent with what is commonly perceived as mathematical skill, fitting with the framework of the Italian national curricula and the theoretical framework of the INVALSI large-scale assessment, hence perfectly admissible in a test assessing mathematical learning in the context. Nevertheless, they may still present features which are not completely fitting with the prediction of the model (even though their behaviour does not affect the overall reliability of the measurement).

We focused our analysis on a group of over-discriminating items. Via a graphical inspection of the characteristic curves and the distractor plots, on one hand, and a qualitative analysis of the tasks, on the other hand, we highlighted a set of similar features of these items, both in their input (formulation, relationship with didactic praxis, etc.) and in their outputs (missing answers, guessing, etc.).

This mixed approach allowed us to formulate a feasible interpretation of the misfit based on a conjectured behaviour of the students related to the specific features of the items. These interpretations should be verified through an ad hoc qualitative experimental apparatus, based on a direct analysis of students' behaviour.

Our methodology, based on the integration of quantitative evidence given by Rasch analysis and the interpretation from a didactic point of view, suggests a new utilization of Rasch's model also in the case of misfitting items. The deviation of empirical data from the expected theoretical line is a signal of a disturbance factor. Nevertheless, in our case this deviation and these factors can be explained and framed in a coherent setting. Hence, deviations from Rasch's expectations (i.e. misfits) that do not cause concern from a psychometric point of view (i.e. they do not cause unfair estimations) can be conceived as a further "result" of the Rasch model because they may produce relevant information about students' behaviour, allowing us to formulate some specific hypotheses. From a symmetric perspective, these items allow us to better outline the actual construct measured by the test.

Acknowledgements

We thank INVALSI for the data sets allowing our analyses. The views expressed in this work are attributable exclusively to the authors and do not in any way involve the responsibility of INVALSI. In citing the work, it is not appropriate to attribute the arguments expressed herein to INVALSI or its management.

Disclosure statement

The authors declare that they have no potential conflict of interest to disclose.

References

- Alagumalai, S., & Curtis, D. D. (2005). Classical test theory. In S. Alagumalai, D. D. Curtis, & N. Hungi (Eds.), *Applied Rasch Measurement: A book of examples* (pp. 1–14). Dordrecht, The Netherlands: Springer.
- Barbaranelli, C., & Natali, E. (2005). *I test psicologici: Teorie e modelli psicometrici*. Roma: Carocci.
- D'Amore, B. (2014). *Il problema di matematica nella pratica didattica*. Modena: Digital Docet.
- De Lange, J. (2007). Large-scale assessment and mathematics education. In F. K. Lester, Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 1111–1142). Charlotte, NC: Information Age Publishing.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum associates Publishers.
- European Parliament and Council. (2006). Recommendation of the European Parliament and of the Council of 18 December 2006 on key competences for lifelong learning (2006/962/EC). *Official Journal of the European Union*, 394(1), 10–18. Retrieved from <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L:2006:394:FULL&from=EN>
- Feagin, J., Orum, A., & Sjoberg, G. (1991). *A case for the case study*. UNC Press Books.
- Gustafson, J. (1980). Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology*, 33(2), 205–233.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijho.
- INVALSI. (2015). *Rilevazioni nazionali degli apprendimenti 2014–2015: Rapporto tecnico*. Roma: INVALSI.
- INVALSI. (2016). *Rilevazioni nazionali degli apprendimenti 2015–2016: Rapporto tecnico*. Roma: INVALSI.
- INVALSI. (2017). *Rilevazioni nazionali degli apprendimenti 2016–2017: Rapporto tecnico*. Roma: INVALSI.
- Johansson, R. (2003). Case study methodology: A keynote speech at the International Conference *Methodologies in Housing Research*. Royal Institute of Technology in cooperation with the International Association of People–Environment Studies, Stockholm, 22–24 September 2003, 1–14.
- Le, L. (2012). *Item point-biserial discrimination*. Retrieved from <https://www.acer.org/files/Conquest-Notes-5-ItemPointBiserialDiscrimination.pdf>
- Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, 16(2), 878.
- Martin-Löf, P. (1974). Notion of redundancy and its use as a quantitative measure of the discrepancy between a statistical hypothesis and a set of observational data.

- Scandinavian Journal of Statistic*, 1(1), 3–18.
- McDonald, R. (1981). The dimensionality of tests and items. *British Journal of mathematical and statistical Psychology*, 34(1), 100–117.
- Meinck, S., Neuschmidt, O., & Taneva, M. (2017). Workshop theme: “Use of educational large-scale assessment data for research on mathematics didactics”. In G. Kaiser (Ed.), *Proceedings of the 13th International Congress on Mathematical Education: ICME-13* (pp. 741–742). Cham: Springer International Publishing.
- Millman, J., & Green, J. (1989). The specification and development of tests of achievement and ability. In R. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 335–366). New York: Macmillan.
- Molenaar, I., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika*, 55(1), 75–106.
- Morgan, D. L. (1998). Practical strategies for combining qualitative and quantitative methods: Applications to health research. *Qualitative health research*, 8(3), 362–376.
- OECD (2013). *Assessment and analytical framework: Mathematics, reading, science*. OECD Publishing.
Retrieved from <http://dx.doi.org/10.1787/9789264190511-en>
- Primi, C. (2017). Solid findings in mathematics education: A psychometric approach. *CERME 10*, Feb 2017, Dublin, Ireland. Retrieved from <https://hal.archives-ouvertes.fr/hal-01849611/document>
- Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Education Research.
- Smith, R. (1993). Guessing and the Rasch model. *Rasch measurement transactions*, 6(4), 262–263.
- Wright, B. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14(2), 97–116.
- Wright, B. D., Linacre, J. M., Gustafson, J. E., & Martin-Löf, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8(3), 370. Retrieved from <http://www.rasch.org/rmt/rmt83b.htm>
- Wu, M., & Adams, R. J. (2013). Properties of Rasch residual fit statistics. *Journal of applied measurement*, 14(4), 339–355.