Investigating foreign students’ disadvantage in mathematics: A mixed method analysis to identify features of items favouring native students

Investigating foreign students' disadvantage in mathematics: a mixed method analysis to identify features of items favouring native students

Clelia Cascella[a], Chiara Giberti[b]* and Matteo Viale[c]

[a] *School of Environment, Education and Development, University of Manchester, Manchester, UK;*

[b]**Department of Human and Social Science, University of Bergamo, Bergamo, Italy*

[c]*Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy*

Differences in performance of native and foreign students in reading comprehension and mathematics emerge in all standardised assessments, both at national and international levels. This study is the continuation of a previous research project in which we examined citizenship-based differences in Italian mathematics standardised assessments, both at item level and in connection with students' competence in text comprehension. We present a deeper analysis of items which presented a significant Differential Item Functioning (DIF) in favour of native students, and interpret these differences in terms of linguistic difficulties. On the basis of this analysis, we then present a classification of possible item characteristics causing disadvantage to foreign students, going one step further classification of items in terms of reading demand. The results obtained are then supported by further analysis of items similar to those considered in the first part of the present study, but administered in different years: all the results in terms of DIF analysis are confirmed also in the similar items' analysis.

Keywords: citizenship; mathematics differential item functioning; standardised assessment; Rasch; pseudo-longitudinal design

1. Introduction

The issue of equity in mathematics education has been widely discussed in recent decades, and international standardised assessments (such as PISA and TIMSS) have highlighted strong differences between students regarding their social and economic backgrounds, gender and other factors (OECD, 2015). Inequalities between students still exist both in developed and developing countries, and educational policies in this field should thus be rethought because, as highlighted by Ongaki and Musa:

> Equity in education is more than an issue of fairness and distributive justice, especially in the current period when many countries are trying to develop their human resources as one element in enhancing growth and international competitiveness in the job market. Unequal education implies that human potential is being wasted, and that some individuals do not have the competence to perform well in a modern society (Wößmann & Schutz, 2006; Ongaki & Musa, 2014)

From this perspective, it is important to analyse and interpret these differences in order to understand how to provide all students with the possibility to exploit their full potential in the school system and also, later, in society.

In this paper, we will analyse differences in mathematics performance between students from immigrant backgrounds and native students. There are many factors that might explain an inferior performance of immigrant students in mathematics[1]: a more complex and difficult background (e.g. Entorf & Lauk, 2008; Giannelli & Rapallini, 2016), generational status and age of arrival in the destination country (Schleicher, 2006; Böhlmark, 2008) and linguistic barriers (e.g. Carhill et al.,

[1] In accordance with the OECD and INVALSI procedure, we categorised students by citizenship status using the following criteria: Native (i.e., student born in Italy with at least one parent born in Italy), first-generation foreign/immigrant (i.e., student not born in Italy to parents not born in Italy), and second-generation foreign (i.e., student born in Italy to parents not born in Italy).

2008) are all factors that can contribute to their greater difficulties in the school context and, more specifically, in mathematics (Morley, Leach, & Lugg, 2009). We decided to focus on one specific factor, i.e., linguistic barriers, and analyse in greater depth the relationship between student performance in mathematics and in reading comprehension. Our research is based on Italian standardised assessments, namely the INVALSI tests (promoted by the Italian National Institute for the Evaluation of Educational Systems), which are administered every year at grades 2-5-6[2]-8-10 and 13. INVALSI administers two different tests in all grades: one to examine students' competence in mathematics and one to examine their competence in the Italian language (grammar and reading comprehension)[3]. In a recent work (Cascella & Giberti, 2020), we used a quantitative statistical approach to analyse the association between students' competence in mathematics and in reading comprehension, also considering their background (immigrant or native). Furthermore, we identified items showing a statistically significant DIF between native and immigrant students (first and second generation). Among items showing a statistically significant DIF, we focused on those showing a moderate to large DIF (Zwick, 2012; Zwick, Thayer & Lewis, 1999): the magnitude of detected DIF determines whether "the effect of that DIF is of substantive importance" (Adams & Wu, 2010, p. 5).

In this paper, we first analyse these items in terms of reading demand and propose a new classification based on linguistic and formulation features which can explain the differences that emerge between native and immigrant students' performance in mathematics.

Then, we analyse all the data collected by INVALSI in each grade between 2010 and 2017 to identify items showing characteristics similar to those analysed in the first part of the present study; this is carried out to test our hypothesis that the specific linguistic characteristics of these items are systematically associated with the disadvantage of foreign students compared with native students.

[2] Grade 6 tests were administered until 2013.

[3] Since 2018, INVALSI also administers a test at grades 5, 8, 10 and 13 to evaluate English language competence.

2. Immigrant students in the Italian context

In 2019, the proportion of students with an immigrant background in Italian schools was approximately 12% in primary school, 10% in lower secondary school and 7% in upper secondary schools (source: ISTAT - Italian National Institute of Statistics). National and international standardised assessments highlight how native students outperform foreign students in mathematics in all grades. For example, PISA surveys showed that although this gap is decreasing in some countries (OECD, 2013), Italy is one of the countries in which the gap is greater and still growing (OECD, 2016; Giberti & Viale, 2019). Interesting findings emerged from the last TIMSS (Trends in International Mathematics and Science Study) surveys administered in 2019 at grade 4 and 8, in terms of results in mathematics as compared with the frequency that students speak the language of the test at home (Mullis et al., 2020). In Italy, approximately 75% of the students in both grades stated that they always speak Italian at home, and the results in mathematics of these students are higher - 17 points higher at grade 4 and 14 points at grade 8 - than the international average (always referring to students speaking the language of the test at home). On the other hand, if we consider students who rarely speak the language of the test at home (in Italy, 11% at grade 4 and 7% at grade 8), we find that their results are lower than, or similar to, the results of the same category of students at an international level, both at grade 4 and grade 8. The results of these students are, obviously, also significantly lower than the results of students who always speak Italian at home (gap of 23 points at grade 4 and 41 points at grade 8).

Considering INVALSI tests administered in the same year (INVALSI, 2019), we observe that native students outperform immigrant students in both mathematics and Italian tests, and this gap is even larger for first generation foreign students. The gap in mathematics seems to be greater in the first cycle of education (primary and lower secondary school) than in upper secondary school, as shown in the following graphs (Fig. 1).

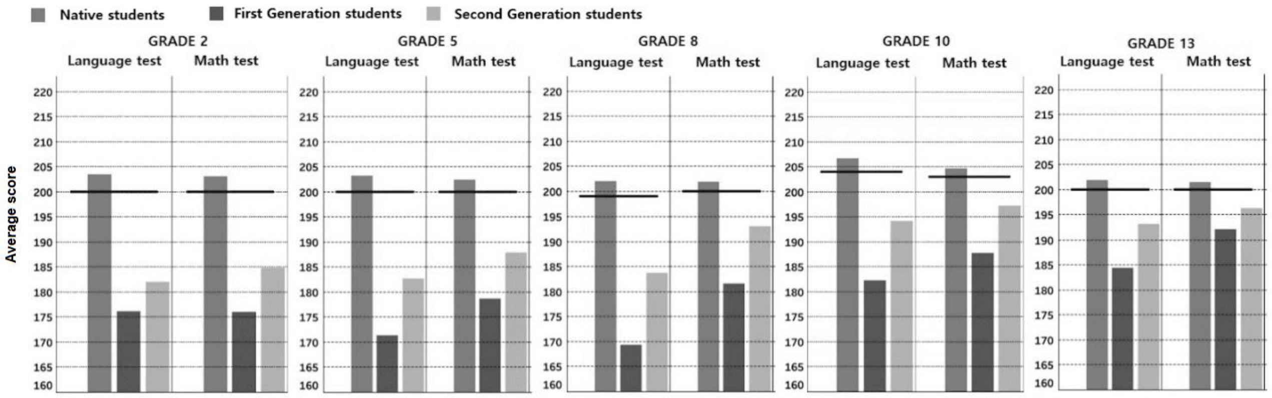


Figure 1. INVALSI 2019 results for native students and foreign students (first and second generation). Source: INVALSI Report (INVALSI, 2019). The distribution of scores has been linearly transformed so that the score average (for each test) equals 200 (st. dev. = 40).

3. Language-related difficulties in mathematics

Within the scholastic tradition, mathematics and Italian language have often been perceived as separate disciplines and rarely interact. Although there have been illustrious research studies since the 1970s (for the Italian context, see, among others: Altieri Biagi, 1978; Altieri Biagi, Pasquini & Speranza, 1979; Altieri Biagi & Speranza, 1981; Altieri Biagi et al., 1982), nowadays there is a growing awareness of the need for shared efforts between Italian and mathematics teachers on the language front, trying to guide students to a full appropriation of the mathematics text.

It is increasingly evident that success and failure in mathematics are linked to the ability to correctly decode the texts through which school mathematics is presented, such as the theoretical explanations in textbooks and the texts of problems and exercises proposed (Viale, 2019; Ferrari, 2021; Radford & Barwell, 2016). However, school mathematical texts, exercises and (above all) problems, present many difficulties due to their linguistic peculiarities and the contrast between the use of language in everyday experience and that used in disciplinary texts. Difficulties can lurk in different linguistic planes (Lavinio, 2007) and vary according to the language used, as highlighted by Bergqvist, Theens and Österholm (2018) in a recent work based on 83 PISA tasks translated into three different languages (German, Swedish and English).

Considering Italian language, from a textual point of view, the texts used in mathematics (both in standardized tests and in daily practices, in the framework of practices that go beyond a specific language) are by nature non-continuous or mixed, with a continuous interweaving of text, images, formulas, etc. that requires a specific cognitive effort (Lavinio, 2007; Demartini & Sbaragli 2021). From a syntactic point of view, despite some exceptions, a simple minimalist syntax normally prevails in school mathematics texts, with short sentences and a preference for coordination. However, there is a great deal of implicit subordination, especially in the gerund (as in the stereotypical "Sapendo che…" - "knowing that…"), passive forms, peculiar uses of the subjunctive ("Sia B il lato…"[4]) and recourse to nominalisation, i.e., expressing actions with nouns instead of verbs ("Si proceda alla sottrazione di X alla somma ottenuta" instead of "Sottrai 3 alla somma ottenuta" - "Proceed with the subtraction of X from the resulting sum" instead of "Subtract X from the sum obtained"). All these phenomena contribute to increasing linguistic complexity of texts because of their remoteness from the everyday language (inside and outside school) that is familiar to the students.

Considering the lexicon, school mathematics alternates between a use of language that is not particularly sophisticated but as simple and as appealing as possible, and the necessary recourse to a gradually increasing number of technical terms that are interconnected. It is precisely this dialectic, fluctuating between common everyday and specific technical vocabulary, which gives rise to misconceptions with the result of constant difficulty (D'Amore, 2000).

[4] Even in Italian this form is unusual and it is difficult to translate it into English. The fixed expression "Sia B il lato…", which is very widespread in the Italian didactic tradition of teaching mathematics, is based on a use of the subjunctive that is obsolete in today's Italian. One possible English translation could be "Let b be the side...", but this choice does not truly render the idea of how the mathematical usage of the words differs greatly from the natural language meaning.

All these aspects contribute to increasing the linguistic complexity of mathematics texts (both in theory chapters and exercises) and sometimes make the student's approach to textbooks problematic without a teacher’s support (Sbaragli & Demartini, 2021). A systematic analysis of the linguistic difficulties encountered by students in dealing with mathematical texts suggests that mathematical competence is closely interrelated with literacy competence, and that a substantial proportion of difficulties in mathematics can also be traced to a problem of text comprehension.

A recent research study on the Italian situation, based on PISA 2012 data, was conducted by Ajello, Caponera and Palmerio (2018) and confirmed the strong link between reading competency and mathematics performance. In this work, PISA items were classified in terms of reading demand, following a specific framework in line with that proposed by Mullis and colleagues in TIMSS reports (Mullis, Martin & Foy, 2013). Furthermore, starting from the results of native Italian students involved in the TIMSS and PIRLS (Progress in International Reading Literacy Study) surveys in 2011, Caponera and colleagues (2016) confirmed the influence of reading literacy on mathematics achievement, while they found that the correlation between students’ competencies in mathematics and reading in Italy was weaker than in other countries. Following the same criteria used by Ajello and colleagues, they considered items with higher and lower reading demand, and analysed students’ performance in relation to their reading ability: the level of reading demand of the task seems not to affect answers of students with high reading ability; on the contrary, students with low reading ability showed differences in their performance according to the cognitive demand in reading the tasks.

In a recent paper (Cascella & Giberti, 2020), we analysed large-scale assessment data collected by the INVALSI to understand if - and, possibly, how - the relationship between students’ reading competences and their performance in mathematics changes by age (from primary to lower secondary education). Results showed that (i) ‘reading demand’ is just one of the possible factors explaining differences in mathematics performance among students categorised by citizenship; and, that (ii) foreign students are not always disadvantaged as compared with native students in mathematics. The results of the previous research study are reported in the Appendix 1 (Cascella & Giberti, 2020).

4. Research questions

In line with previous studies, this paper aims to explore the possible association between students' reading skills and their performance in mathematics. It also aims to analyse items showing a statistically significant and large DIF (Zwick, 2012) between 'native' and 'foreign' students. We hypothesised that the formulation of an item and, in particular, the use of specific words and/or syntactic structures may be difficult for students operating in a second language, and thus explain differential item functioning as a consequence of their citizenship.

More precisely, this paper aims to answer the following research questions:

1) which features of items showing a significant and moderate to large DIF can explain a citizenship gap?
2) after identifying those items with similar syntactic and linguistic characteristics, is the DIF between native and foreign students confirmed also in other tests administered by INVALSI to students belonging to different populations?

Compared to our previous study, this paper thus goes a step further as it sets out to understand the mechanisms that cause Differential Item Functioning between native and foreign students, thus contributing to the debate about the complex relationship between students' reading comprehension skills and their performance in mathematics.

5. Methodology

According to the classification adopted by Creswell and Plano-Clark (2017), the design can be considered a multilevel sequential explanatory mixed methods research design. In fact, neither a blind quantitative datum (however broad it may be) nor a specific and necessarily contextualised focused study (however in-depth it may be) can in itself offer a sound and convincing account of a complex phenomenon like that under consideration, whose evidence emerges at systemic level. This implies a gradual switching from quantitative large-scale evidence to qualitative analysis with in-depth observation that will be confirmed by further quantitative analysis.

This mixed method research has been adopted in several recent research in math education also with the aim of quantifying solid finding of mathematics education (Bolondi & Ferretti, 2021) and to understand the causes of differences between students performances such as gender gap (Ferretti & Giberti, 2021; Cascella, Giberti & Bolondi, 2020).

In this paper, we build on results from the quantitative analysis presented in a previous work (Cascella & Giberti, 2020) to identify item features that we believe useful in explaining the Differential Item Functioning which emerges in favour of native students.

Results presented in the current paper are thus from a two-step methodology:

- STEP 1. Qualitative analysis: Reading demand and item features

  In this step, we considered the mathematics items showing a statistically significant and moderate/large DIF in favour of native students in the INVALSI tests considered in the previous research (Cascella & Giberti, 2020). Then, we quantified the reading demand of the selected items and, using a qualitative analysis (described in section 5.3.1), we formulated new criteria that might explain in more detail the differences between native and immigrant students' performance in mathematics;

- STEP 2. Quantitative analysis: Differential Item Functioning

  In the second step, we analysed all the data collected by INVALSI between 2010 and 2017 in an attempt to find items similar to those identified in Step 1. We define as 'similar' two or more items focused on the same mathematical content, with the same question intent (as specified by INVALSI) and with a similar formulation: thus, approximately the same number of words, the use of the same (or similar) table/figure to represent data and/or the same layout (i.e., a text followed by a figure/table, or vice versa). Then, we analysed these items following the new criteria developed in Step 1, and analysed their DIF by citizenship to confirm the hypothesis that specific linguistic characteristics are systematically associated with a position of disadvantage held by foreign students as compared with native students, irrespective of the sample or population they are from.

5.1 Data

INVALSI administers both open-ended and multiple-choice mathematics items. All students' answers have been dichotomised (0 = wrong answer; 1 = correct answer) before analysing them within the framework of Rasch analysis (Rasch, 1960/1980).

The Rasch model is particularly suitable to pursue the aims of the present study as it postulates that (i) the probability of answering an item correctly depends on students' relative ability, that is their ability compared with item difficulty; and, (ii) no other factors (including characteristics like gender, citizenship and so on) can affect such a probability.

Within the framework of the Rasch analysis, both students and items are scaled along the same latent trait (which, in the current study, is the students' ability in mathematics), thus allowing comparability between sub-groups of students, sub-groups of items, and (sub-groups of) students with (sub-groups of) items. Comparing groups of students - typically matched on ability (i.e., showing relatively similar levels of mathematics competence) - is the idea underlying Differential Item Functioning analysis (Osterlind & Everson, 2009). DIF refers to each single item behaviour in sub-groups of students matched on ability and clustered by one personal student attribute (in this case, citizenship). DIF is employed here to analyse INVALSI data, and aims at understanding whether (and if so, to what extent) item difficulty (and thus students' ability in mathematics) is affected by citizenship.

DIF analysis was carried out in RUMM2030 (Andrich, Sheridan & Luo, 1997-2012). For each item, we plotted its Item Characteristic Curve (ICC), one for native and one for first- and/or second-generation foreign students. ICCs express the probability of students (spanning a wide range of mathematical ability levels) answering each item correctly. Therefore, by comparing ICCs plotted by citizenship, we showed how the probability of tackling each item successfully varies among students who are matched on ability but have different citizenship status.

Within such an analytical framework, any Differential Item Functioning (e.g., by citizenship) is to be considered a violation of the Rasch model's assumptions. Nonetheless, when such a violation falls within certain tolerance intervals (Wright & Linacre, 2004), it is not disruptive to measurement but, on the contrary, can be considered as informative from a substantive point of view (Bolondi & Cascella, 2020).

5.1.1. STEP 1: Reading demand and item features.

The data considered in the first step are the same as in the previous study (Cascella & Giberti, 2020). Following a pseudo-longitudinal approach, we analysed three INVALSI mathematics tests: grade 5 test administered in 2009, grade 6 in 2010, and grade 8 in 2012.

The INVALSI sample, representative of the whole population is described, for each grade, in Table 1; not all the information provided for grade 6 and 8 are available for grade 5.

Table 1

INVALSI data considered in Step 1 and sample features.

	Grade 5 2008/09	Grade 6 2009/10	Grade 8 2011/12
Number of students	43585 (100%)	41635 (100%)	25556 (100%)
Gender			
- Boys	-	20672 (50%)	12932 (51%)
- Girls	-	19307 (46%)	12624 (49%)
Citizenship status			

- Native	34713 (80%)	37102 (89%)	23070 (90%)
- Foreign students (no classification by generational status)	3402 (8%)		
- First-generation foreign students	-	2694 (6%)	1694 (7%)
- Second generation foreign students	-	1560 (4%)	792 (3%)
Regularity			
- Regular	-	36680 (88%)	22310 (87%)
- In advance	-	479 (1%)	311 (1%)
- Retained	-	2949 (7%)	2935 (11%)

5.1.2. STEP 2: Differential item functioning

In the second step of this study, we considered the sample data collected by INVALSI at grade 5, 6, and 8 from 2010 to 2017 (Table 2). From the items of these 20 INVALSI tests, we identified items as in the previous step, and carried out a more in-depth qualitative analysis of their lexical and syntactic characteristics, followed by a DIF analysis.

Table 2

INVALSI data considered in Step 2 and sample features.

	2009/10	2010/11	2011/12	2012/13	2013/14	2014/15	2015/16	2016/17
GRADE 5								
Number of students	35566 (100%)	31563 (100%)	30869 (100%)	24773 (100%)	25348 (100%)	22030 (100%)	25282 (100%)	25482 (100%)
Gender								
- Boys	17628 (50%)	15961 (51%)	15453 (50%)	12456 (50%)	12778 (50%)	11252 (51%)	12984 (51%)	12838 (50%)
- Girls	16912 (48%)	15598 (49%)	15415 (50%)	12297 (50%)	12522 (49%)	10773 (49%)	12296 (49%)	12601 (49%)
Citizenship status								
- Italians	32038 (90%)	27875 (88%)	27599 (89%)	22057 (89%)	22769 (90%)	19800 (90%)	22759 (90%)	22218 (87%)
- First-generation foreign students	1790 (5%)	1491 (5%)	1413 (5%)	1074 (4%)	868 (3%)	715 (3%)	704 (3%)	567 (2%)

- Second-generation foreign students	1492 (4%)	1436 (5%)	1742 (6%)	1559 (6%)	1628 (6%)	1470 (7%)	1790 (7%)	1740 (7%)
Regularity								
- Regular	33299 (94%)	29872 (95%)	29274 (95%)	23614 (95%)	24269 (96%)	21229 (96%)	24403 (97%)	24468 (96%)
- In advance	333 (1%)	664 (2%)	445 (1%)	329 (1%)	352 (1%)	248 (1%)	311 (1%)	409 (2%)
- Retained	988 (3%)	1025 (3%)	1150 (4%)	793 (3%)	679 (3%)	10420 (4%)	566 (2%)	561 (2%)
GRADE 6								
Number of students	41635 (100%)	40651 (100%)	39668 (100%)	27504 (100%)				
Gender								
- Boys	20672 (50%)	20732 (51%)	20207 (51%)	13914 (51%)				
- Girls	19307 (46%)	19919 (49%)	19460 (49%)	13542 (49%)				
Citizenship status								
- Native	37102 (89%)	36180 (89%)	35204 (89%)	24274 (88%)				
- First-generation foreign students	2694 (6%)	2439 (6%)	2269 (6%)	1524 (6%)				
- Second-generation foreign students	1560 (4%)	2032 (5%)	1999 (5%)	1593 (6%)				
Regularity								
- Regular	36680 (88%)	36179 (89%)	35863 (90%)	25098 (91%)				
- In advance	479 (1%)	610 (2%)	834 (2%)	307 (1%)				
- Retained	2949 (7%)	3862 (10%)	2969 (7%)	2039 (7%)				
GRADE 8								
Number of students	25626 (100%)	25558 (100%)	25556 (100%)	28153 (100%)	28108 (100%)	28494 (100%)	27955 (100%)	28051 (100%)
Gender								
- Boys	12894 (50%)	13034 (51%)	12932 (51%)	14221 (51%)	13877 (49%)	14524 (51%)	14165 (51%)	14219 (51%)
- Girls	12732 (50%)	12524 (49%)	12624 (49%)	13868 (49%)	14077 (50%)	13959 (49%)	13759 (49%)	13826 (49%)
Citizenship status								
- Native	23720 (93%)	23002 (90%)	23070 (90%)	25111 (89%)	25076 (89%)	25608 (90%)	25148 (90%)	24940 (89%)

- Foreign students (no classification by generational status)	1906 (7%)	2556 (10%)						
- First-generation foreign students			1694 (7%)	1802 (6%)	1544 (5%)	1481 (5%)	1244 (4%)	874 (3%)
- Second generation foreign students			792 (3%)	1116 (4%)	1301 (5%)	1369 (5%)	1500 (5%)	1472 (5%)
Regularity								
- Regular	22492 (88%)	22363 (87%)	22310 (87%)	24729 (88%)	24883 (89%)	25295 (89%)	25295 (90%)	25547 (91%)
- In advance	280 (1%)	256 (1%)	311 (1%)	336 (1%)	478 (2%)	454 (2%)	284 (1%)	365 (1%)
- Retained	2854 (11%)	2939 (11%)	2935 (11%)	2996 (11%)	2586 (9%)	2731 (10%)	2342 (8%)	2128 (8%)

All the INVALSI samples are equally distributed by gender. The number of both first- and second-generation foreign students is significantly lower than that of native students but the model of choice in this study, the Rasch model (1960/1980), guarantees invariance of measurement across sub-groups of students, regardless of number.

5.3. Analytical strategy

In Cascella and Giberti (2020), we considered the grade 5 test administered in 2009, grade 6 in 2010, and grade eight in 2012. The results showed that: (i) at grade 5, 5 items (D1, D8d, D22, D25a, D26) showed a statistically significant DIF between native and foreign students, in favour of the first category (except in item D22); (ii) at grade 6, 8 items (D6, D8a, D8b, D9, D13, D16, D27b) showed a statistically significant DIF in favour of natives; and, (iii) at grade 8, 5 items (D10d, D11b, D14, D19, D22) in favour of first-generation foreign students.

In addition to statistical significance, we also analysed DIF magnitude (or size) given by the difference (in absolute terms) between the difficulty parameter based on answers provided by native students and that based on the answers given by first-generation foreign students. Zwick (2012) classified DIF magnitude in three groups: negligible, when DIF magnitude is lower than 0.43;

moderate, when DIF magnitude ranges between 0.43 and 1; and high, when DIF magnitude is above 1.

At grade 5, no items showed a DIF magnitude greater than 0.43; this might also be due to the fact that in this specific case, the data did not differentiate between first-generation immigrant students and second-generation immigrant students, thus the presence of second-generation students might mitigate the magnitude of the DIF detected. Considering grade 6 tests administered in 2010, the items D6, D8a and D27b, and D10d showed a DIF magnitude greater than 0.43. Item D10d showed a differential item functioning in favour of first-generation foreign students, while the remaining items all favoured native students.

Finally, at grade 8, 10 items showed a significant DIF between native and first-generation students: 5 in favour of natives (D1, D3b, D4b, D9a, D24) and 5 in favour of foreign students (D2b, D6, D10b, D11, D19b). Items showing a DIF magnitude greater than 0.43 at grade 8 were D1 and D24 in favour of the native students, and D6 in favour of first-generation foreign students.

Similar results were found when we considered the differences between native and second-generation students both in grade 6 and in grade 8, but DIF magnitude was lower and, sometimes, statistically insignificant. In Appendix 1, we present results of DIF analysis by citizenship as reported in the previous study.

In line with Cascella and Giberti (2020), we focused on the items showing a statistically significant DIF (Zwick, 2012). In particular, to investigate foreign students' difficulties in mathematics, we focused on the items showing DIF in favour of the Italian students. Furthermore, we decided to exclude true/false items from this analysis because of the high possibility of responding correctly to these questions simply by answering randomly (only one true/false of the items highlighted DIF). The following table (Table 3) lists the items considered in the current study (items identified with * showed a DIF magnitude greater than 0.43).

Table 3

Items which showed a significant DIF in the previous work (Cascella & Giberti, 2020)

Grade	Items
5	D1 - D26
6	D6* - D8a* - D8b - D9 - D13 - D16 - D27a - D27b*
8	D1* - D3b - D4b - D9a - D24*

For each of the items listed in Table 3, we performed a 2-step analysis, as described here below.

5.3.1. STEP 1: Reading demand and item features.

In the first step of this research study, we analysed and discussed in more depth the characteristics of items showing a statistically significant DIF. Then, via qualitative analysis, we categorised these items according to their features. First of all, following the criteria proposed by Ajello and colleagues (2018), we classified all the items considered in this study on the basis of the reading demand. To classify each item in terms of low, medium or high reading demand, we then considered:

I. number of words in the natural language

II. presence of images (graphs, tables or other images)

III. presence of mathematics symbolic language

IV. presence of specialised vocabulary

Following Ajello and colleagues (2018), items were classified with high reading demand if the number of words was higher than 200 (criteria I) and at least two out of the three other criteria (II, III and IV) were met. Items were classified with low reading demand if the number of words was lower than 100 with just one of the other three criteria met. All the other items were classified as medium reading demand.

Then, on the basis of previous results on citizenship differences in mathematics and reading comprehension (Giberti & Viale, 2017; Giberti & Viale, 2019; Viale, 2019), we identified other possible features (see section 6.1) that might explain the underperformance of foreign students, proposing a second classification of the same items on the basis of this criteria.

5.3.2. STEP 2: Differential item functioning.

To strengthen our hypothesis about the relationship between citizenship gap in mathematics and linguistic characteristics of the items showing a DIF by citizenship, thereby answering the second research question, we compared the items identified in our previous study (Cascella & Giberti, 2020) with other similar items (similar in both their mathematical content and phrasing) which were administered by INVALSI to different students' samples in different academic years. Then, we performed a DIF analysis to explore possible differences in the probability of tackling each of those items successfully depending on students' citizenship status.

We claim that our proposed analytical strategy (based on the comparison of answers given by students from different INVALSI samples to similar mathematical items) can significantly contribute to the advancement of knowledge: even though the Rasch analysis allows the comparison of groups of students matched on some characteristics (such as citizenship) and even though INVALSI samples are statistically representative of the native students' population, comparing results based on different samples can allow us to (i) gather results from just one sample's characteristics, and (ii) identify possible patterns and regularities between the items' characteristics and students' performance (differentiated by citizenship-status), thus confirming or refuting our hypotheses.

6. Results and discussion

6.1. STEP 1: Reading demand and item features

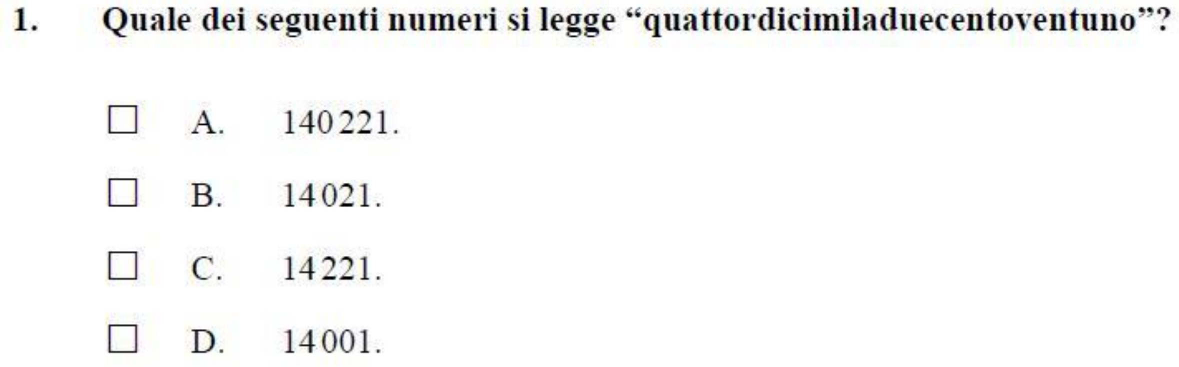
Following the criteria descripted in the methods section proposed by Ajello and colleagues (2018), we identified the reading demand of all the items presenting a significant DIF in favour of Italians. INVALSI mathematics tasks are all designed to minimise the influence of linguistic competence if the question intent of the item is not specifically linked to such ability. For this reason, all the items presented a low number of words: the only item with more than 100 words (yet still less than 200) is item D6 of the grade 6 test; all the other items include fewer than 100 words. All the items were then classified in terms of low or medium reading demand (Table 4), on the basis of the other three criteria (symbols, specialised vocabulary and presence of images) and none of these were classified as high demand

Table 4

Classification of items in terms of reading demand following the criteria suggested by Ajello and colleagues (2018). Items identified with * displayed a DIF magnitude greater than 0.4.

Item	Grade	Number of words (n)	Use of Symbols	Specialised vocabulary (lexicon)	Presence of graphs, tables or other images	Reading Demand
D1	5	n<100	No	no	no	low
D26	5	n<100	Yes	yes	no	medium
D6*	6	100<n<200	Yes	no	yes	medium
D8a* - D8b	6	n<100	No	yes	no	low
D9	6	n<100	No	no	yes	low
D13	6	n<100	No	yes	yes	medium
D16	6	n<100	No	yes	yes	medium
D27a - D27b*	6	n<100	No	no	yes	low
D1*	8	n<100	No	yes	yes	medium
D3b	8	n<100	No	no	no	low
D4b	8	n<100	Yes	yes	yes	medium
D9a	8	n<100	Yes	yes	yes	medium
D24*	8	n<100	No	yes	yes	medium

The classification based on the reading demand is therefore not sufficient to explain the difference in performance between native and foreign students; a paradigmatic example of this is item D1 of grade 5 test (Fig. 2).



1. Quale dei seguenti numeri si legge "quattordicimiladuecentoventuno"?

□ A. 140 221.

□ B. 14 021.

□ C. 14 221.

□ D. 14 001.

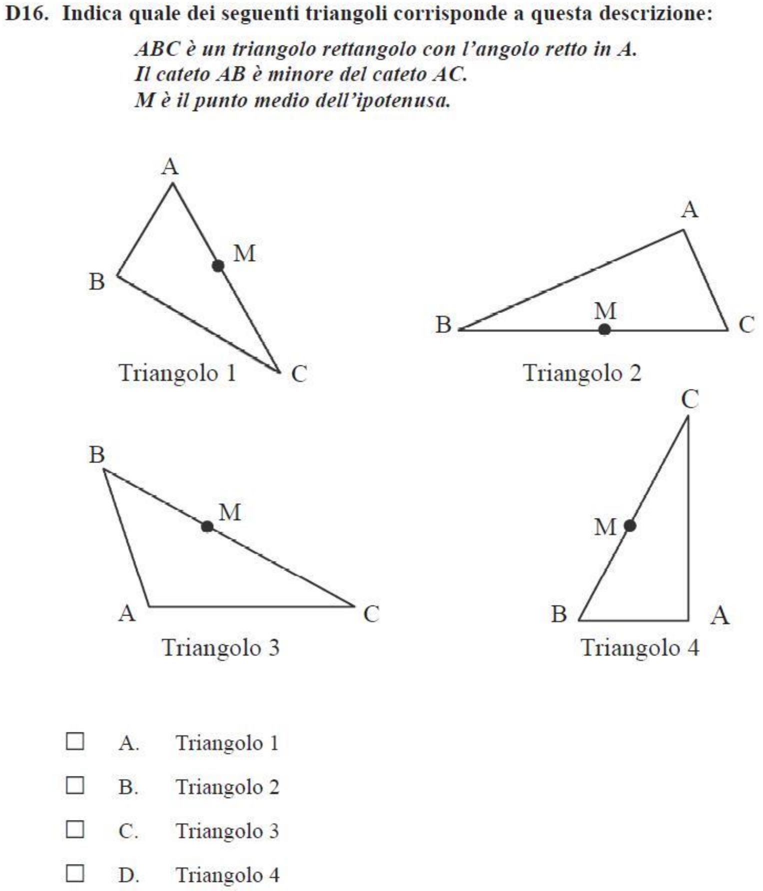
Figure 2. Item D1, grade 5 INVALSI test administered in 2009. (Translation of the authors: *Which of the following numbers is read 'fourteen thousand two hundred and twenty-one'?*)

In this item the number of words is very limited, specialised lexicon is not used (at most, there is the number written in word form) and there is no symbolic language nor images or graphs. Despite this, a significant DIF in favour of native students emerged, which might easily be traced to the difficulty of managing different semiotic registers (natural language and Indian-Arabic numbering system) to represent the same natural number (Duval, 1993).

Furthermore, some of the criteria presented to analyse reading demand might also be useful to explain citizenship differences in mathematics. Indeed, in Table 4 it emerges that 9 of the selected items include a graph, an image or a table. We can therefore classify these items as mixed-text or non-continuous text items. Analyses of data from the INVALSI Italian text comprehension tests show that students have more difficulty with non-continuous or mixed expository texts than with continuous narrative texts. This difficulty can be attributed (among other things) to school practices that favour reading and didactic work using continuous narrative texts. The cognitive effort involved in the joint reading of text and other elements (figures, graphs, formulas) that often recur in mathematics questions may account for some difficulties encountered by students in specific questions and may explain the difference in performance between natives and foreigners (Giberti & Viale, 2019).

A typical example of mixed-text presentation is item D16 administered in the INVALSI grade 6 test in 2010 (Fig. 3): in this case we have the text of the assignment, the text describing the triangle which includes mathematical symbols (ABC, AB, …) and the four pictures; in order to answer correctly, students have to connect all these elements and maintain simultaneous control of different registers and representations.

Certainly, the mixed-text is not the only obstacle for foreign students tackling this item; in this case, also the lexicon used might explain the significant DIF observed.



D16. Indica quale dei seguenti triangoli corrisponde a questa descrizione:

ABC è un triangolo rettangolo con l'angolo retto in A.
Il cateto AB è minore del cateto AC.
M è il punto medio dell'ipotenusa.

A
M
B
C
Triangolo 1
A
B
M
C
Triangolo 2
B
M
A
C
Triangolo 3
C
M
B
A
Triangolo 4

☐ A. Triangolo 1

☐ B. Triangolo 2

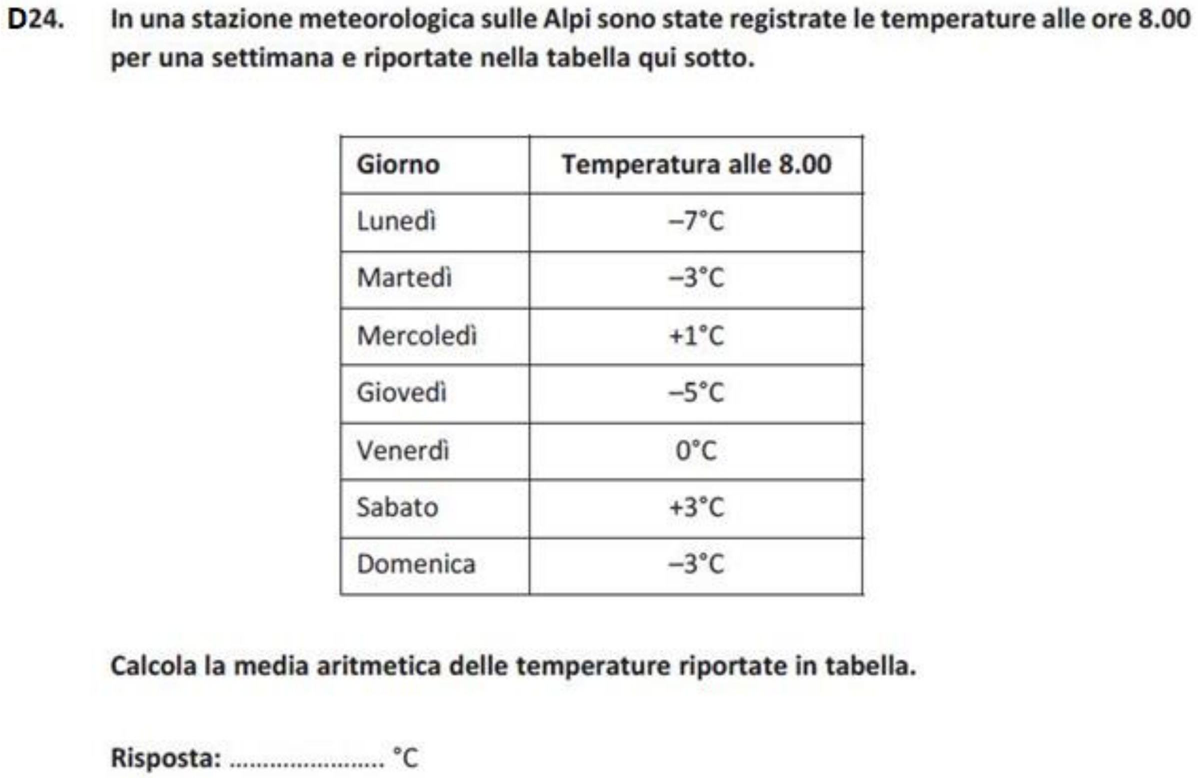
☐ C. Triangolo 3

☐ D. Triangolo 4

Figure 3. Item D16, grade 6 INVALSI test administered in 2010. (Translation of the authors: *Indicate which of the following triangles fits this description: ABC is a right-angled triangle with a right angle at A. The cathetus AB is shorter than the cathetus AC. M is the midpoint of the hypotenuse.*)

Indeed, the presence of specialized lexicon (e.g., 'cateto'/cathetus) might influence foreign students who have not yet achieved complete mastery of the Italian language, and 8 of the selected items meet this criterion. At first glance, lexical complexity can coincide with the word belonging to the 'basic vocabulary' of the Italian language (De Mauro, 1994, 2016; Chiari, 2017), comprising approximately 7,000 words known to all speakers, or to the 'common lexicon', the 14,000 most commonly-used words. However, despite the concrete lexical difficulties that students may encounter in decoding the text of the question, we are led to favour the idea that students are in fact familiar with the lexicon encountered: for example, a specialist term such as ‘ipotenusa’ (hypotenuse) is not among the words most known to general language speakers but is known to students because it is a common term in the school syllabus; on the contrary, words that are part of the fundamental or common lexicon, such as ‘aiuola’ (flowerbed) or ‘acropoli’ (acropolis) may be not very close to students' experience. A further element of difficulty is given by polysemic words with a divergence of meaning between the use of the word in everyday language and the specialist use of the term (e.g., ‘scala’-stair/scale or ‘angolo’-corner/angle).

Another item revealing these two features (mixed-text and lexical complexity), which showed a significant DIF with a magnitude greater than 0.43, is item D24 administered by INVALSI in 2012 at grade 8 (Fig. 4). Indeed, in this item the number of words is really limited but students have to manage both a specific term (‘arithmetic mean’) and a mixed text; in this case, it is not strictly necessary to comprehend the whole text and the situation described: the answer could be given also by considering only the specific term ‘media aritmetica’ (arithmetic mean) and the number reported.



D24. **In una stazione meteorologica sulle Alpi sono state registrate le temperature alle ore 8.00 per una settimana e riportate nella tabella qui sotto.**

Giorno	Temperatura alle 8.00
Lunedì	–7°C
Martedì	–3°C
Mercoledì	+1°C
Giovedì	–5°C
Venerdì	0°C
Sabato	+3°C
Domenica	–3°C

Calcola la media aritmetica delle temperature riportate in tabella.

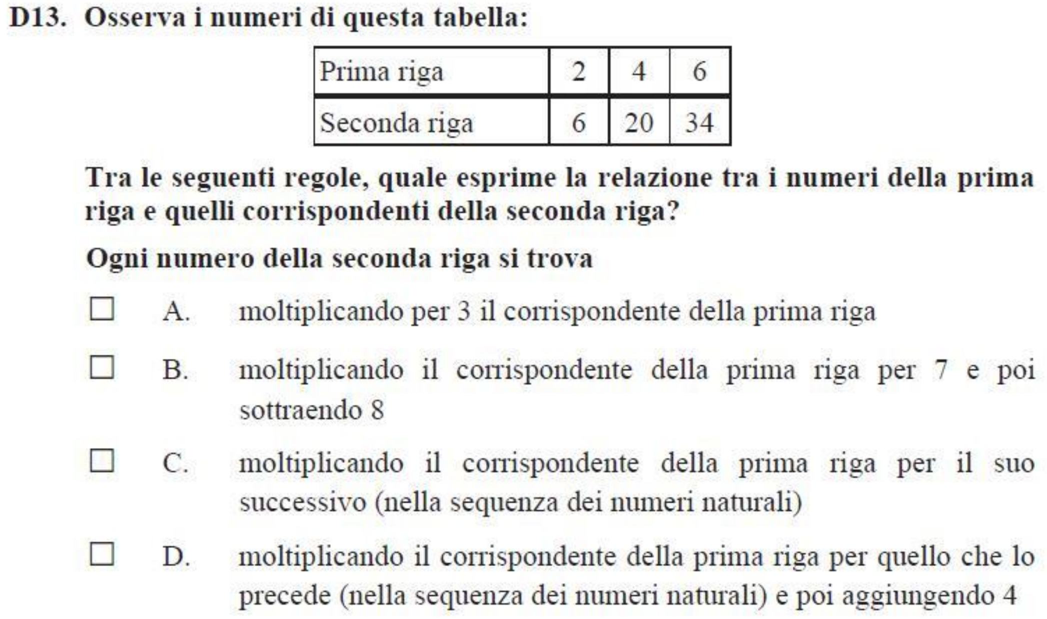
Risposta: .... °C

Figure 4. Item D24, grade 8 INVALSI test administered in 2012. (Translation of the authors: *At a weather*

station in the Alps, temperatures at 8:00 a.m. were recorded for one week and shown in the table below. Calculate the arithmetic mean of the temperatures shown in the table.)

Another characteristic of the formulation of a question to be taken into account when analysing linguistic complexity is syntax. By syntactic complexity we mean, on one hand, the presence of hypothetical structures with extensive use of subordinate sentences, which in most cases corresponds to a high average sentence length; on the other hand, a use of morphosyntactic structures that are less common in everyday language (e.g., passive forms, gerund subordinates, etc.) which increase syntactic density even in texts with a normal or low average sentence length (Sciumbata, 2021; Sutherland & Isherwood, 2016).

A paradigmatic example of an item with a high level of syntactic complexity, due in particular to the repeated use of gerund verbs, can be found in D13 administered in the INVALSI grade 6 test in 2010 (Fig. 5).



D13. Osserva i numeri di questa tabella:

Prima riga	2	4	6
Seconda riga	6	20	34

Tra le seguenti regole, quale esprime la relazione tra i numeri della prima riga e quelli corrispondenti della seconda riga?

Ogni numero della seconda riga si trova

□ A. moltiplicando per 3 il corrispondente della prima riga

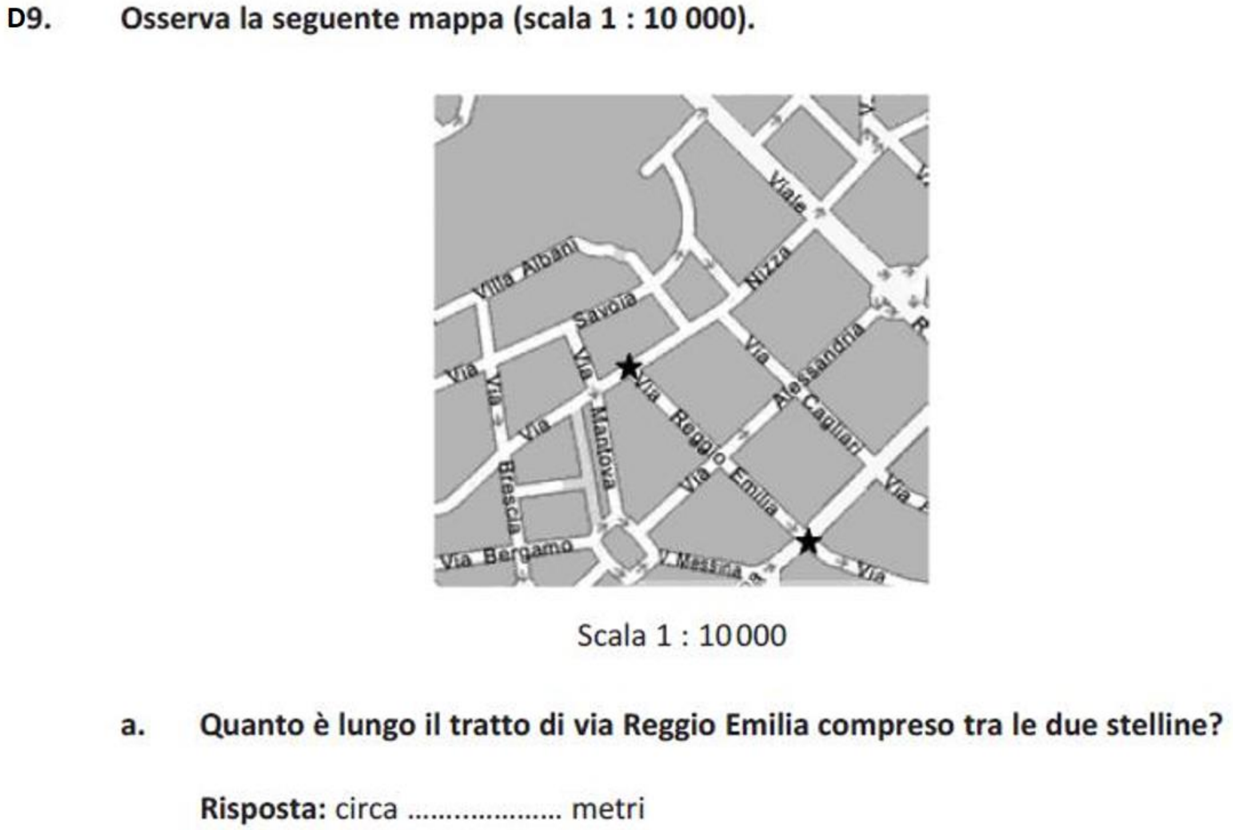
□ B. moltiplicando il corrispondente della prima riga per 7 e poi sottraendo 8

□ C. moltiplicando il corrispondente della prima riga per il suo successivo (nella sequenza dei numeri naturali)

□ D. moltiplicando il corrispondente della prima riga per quello che lo precede (nella sequenza dei numeri naturali) e poi aggiungendo 4

Figure 5. Item D13, grade 6 INVALSI test administered in 2010. (Translation of the authors: *Look at the numbers in this table: First row 2 4 6 Second row 6 20 34 Among the following rules, which one expresses the relationship between the numbers in the first row and the corresponding numbers in the second row? Each number in the second row is obtained by A. multiplying the corresponding number in the first row by 3. B. multiplying the corresponding number in the first row by 7 and then subtracting 8 C. multiplying the correspondent of the first row by its successor (in the sequence of natural numbers) D. multiplying the correspondent of the first row by the one before it (in the sequence of natural numbers) and then adding 4.*)

Finally, following a previous study, we consider that it is not always necessary to read and comprehend the entire text in the question in order to answer correctly (Giberti & Viale, 2019); there are some tasks in which the textual component is very limited and not essential for the interpretation and resolution of the problem, so the question can be easily grasped even without careful reading. If we consider item D9a administered in the INVALSI grade 8 test in 2012 (Fig. 6), we observe that in this case the text is limited and not even necessary: students could understand the situation and request merely by observing the picture and the scale.



D9. **Osserva la seguente mappa (scala 1 : 10 000).**

Villa Albani
Savoia
Viale
Nizza
Via
Via
Via
Via
Mantova
Via Reggio Emilia
Via Alessandria
Via Cagliari
Via Emilia
Brescia
Via Bergamo
V. Messina
Via

Scala 1 : 10000

a. **Quanto è lungo il tratto di via Reggio Emilia compreso tra le due stelline?**

Risposta: circa .... metri

Figure 6. Item D9A, grade 8 INVALSI test administered in 2012. (Translation of the authors: Look at this map (scale 1:10,000). A. How long is the section of Via Reggio Emilia between the two stars? Answer: about……metres)

Despite this fact, this item also revealed a significant disadvantage for foreign students. The explanation of this gap might be related to the use of the polysemic term ‘scala’ (scale and stepladder) and thus to lexical complexity: in a recent work, we interviewed natives and foreign students on a similar task including this term and some foreign students explained that they know the term ‘scala’ only in daily and real situations in the context of “the stepladder that my father uses to change a bulb”.

We then classified (Table 5) all the items, considering the following possible causes of difficulty of foreign students:

I. large amount of text (more than 100 words)

II. non-continuous text or mixed text (alternation between text, figures, graphs, formulas)

III. necessity of text reading

IV. lexical complexity of the text

V. syntactic complexity of the text

VI. type of item (we consider that multiple choice questions help foreign students in answering the question while open-ended and argumentative questions might be an obstacle for students with language difficulties; indeed, in the first case only a receptive linguistic skill is needed, while the second also calls for productive skill)

Table 5

Item classification following our criteria. Items identified with * showed a DIF magnitude greater than 0.4.

Item	Grade	Large amount of text	Mixed text	Necessity of text	Lexical complexity	Syntactic complexity	Reading demand	Type
D1	5	n<100	no	no	yes	no	low	multiple choice
D26	5	n<100	no	yes	yes	no	medium	multiple choice
D6*	6	n>100	yes	yes	no	no	medium	open answer univocal
D8a* - D8b	6	n<100	no	yes	yes	no - medium syntactic difficulty in b	low	open answer univocal
D9	6	n<100	yes	yes	no	yes	low	multiple choice
D13	6	n<100	yes	yes	yes	yes	medium	multiple choice
D16	6	n<100	yes Drawing and definition	yes	yes	no	medium	multiple choice

D27a - D27b*	6	n<100	yes	yes	no	no	low	open answer univocal
D1*	8	n<100	yes	yes	yes	no	medium	multiple choice
D3b	8	n<100	no	yes	no	no	low	open answer, justification
D4b	8	n<100	yes	yes	yes	no	medium	open answer univocal
D9a	8	n<100	yes	no	yes	no	medium	open answer univocal
D24*	8	n<100	yes	no	yes	yes	medium	open answer univocal

The criteria proposed, integrated with that proposed by Ajello and colleagues (2018) to analyse the reading demand, allowed us to explain possible reasons for the existence of a citizenship-based DIF in favour of native students in most of the items. In particular, 11 out of 13 items reveal possible difficulties for foreign students in at least 3 criteria. The DIF in the other two items is easily explained by a lack of language competence: item D1 (Fig. 2) asks students to recognise a high number written in words, and incomplete mastery of the Italian language might prevent students from answering correctly; item D3b asks students to justify their answer and then to write a short explanation.

6.2. STEP 2: Differential item functioning

In this section, we compared the psychometrical functionality of the items presented in the previous paragraph with other items, which were administered in the same grade but in different years and showed the same characteristics in terms of mathematical content (even including the question intent), linguistic characteristics, and layout features.

The following table (Table 6) lists the items considered in this study and items identified as ‘similar’ from other INVALSI tests.

Table 6

Items considered in this study and items identified as “similar” from other INVALSI tests

Item	Grade	Similar Items found in one or more INVALSI tests administered from 2010 to 2017
D1	5	D27 grade 5 2018 D1 grade 5 2016
D26	5	no similar items were found
D6*	6	no similar items were found
D8a* - D8b	6	no similar items were found
D9	6	D12 grade 6 2012
D13	6	no similar items were found
D16	6	D13 grade 6 2013
D27a - D27b*	6	D13 grade 6 2011
D1*	8	D5a grade 8 2013
D3b	8	no similar items were found
D4b	8	no similar items were found
D9a	8	D20a grade 6 2013
D24*	8	D12b grade 10 2015

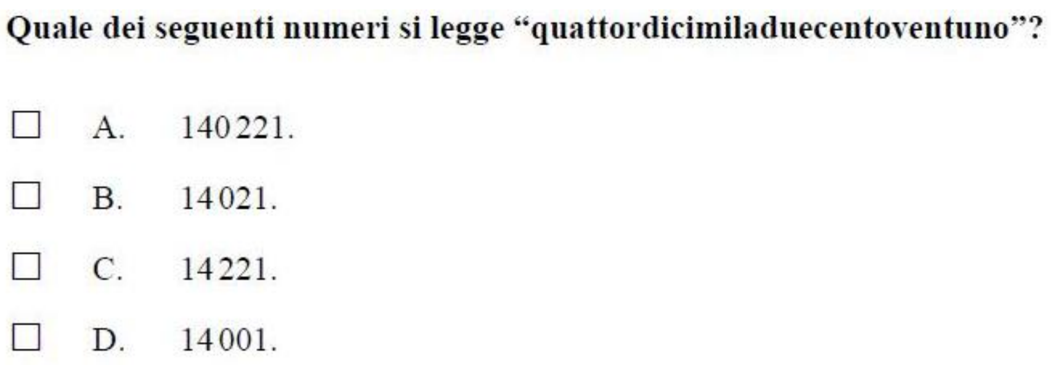
Table 7 and Table 8 report on the comparison between some items analysed in this study (i.e., item D1 administered at grade 5, item D9 at grade 6) and two similar items administered by INVALSI at the same grades but in other years, and thus to different students in different achievement tests (i.e., items D1, D12, respectively).

The items we compared were identical in terms of mathematical content, question intent and formulation. Therefore, the comparison of these items provided information about the relationship between these specific items and foreign students' disadvantage, regardless of the students' sample. In order to compare items' psychometric functionality, the following table reports the items analysed and the comparison between their characteristic curves. In each graph, the probability of answering an item successfully is provided (on the y-axis) according to students' ability (on the x-axis).

Table 7

Comparison between item D1 administered at grade 5 in 2009 and item D1 administered at grade 5 in 2016

Item_D1_Grade_5(2009)



Quale dei seguenti numeri si legge "quattordicimiladuecentoventuno"?

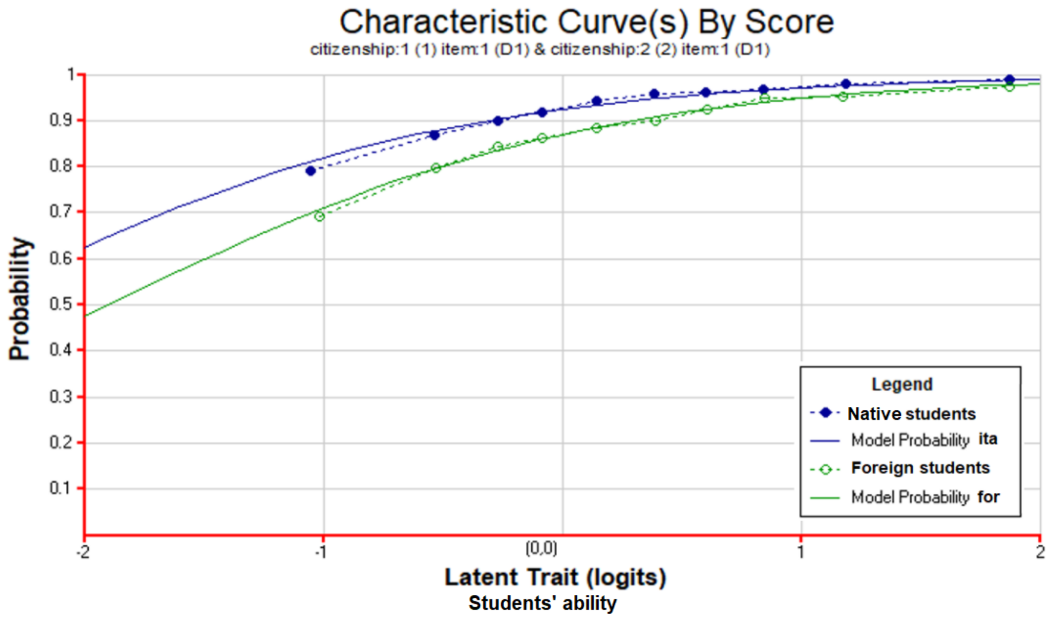
□ A. 140221.

□ B. 14021.

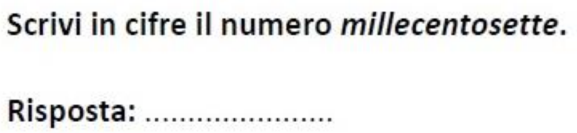
□ C. 14221.

□ D. 14001.

(Translation of the authors: Which of the following numbers is read ‘fourteen thousand two hundred and twenty-one’?)



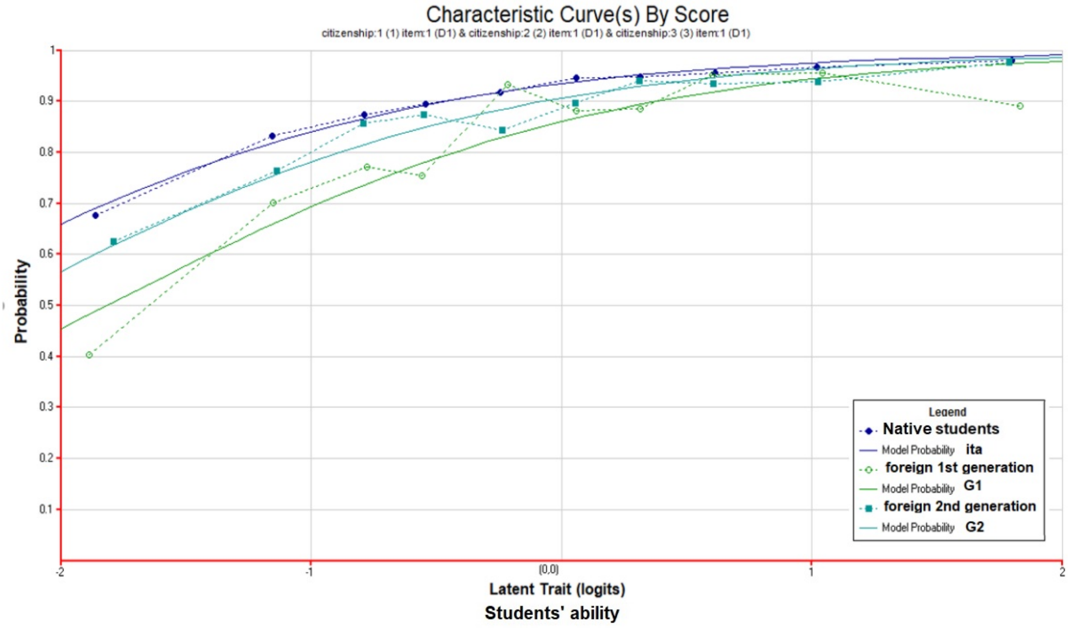
Item_D1_Grade_5(2016)



Scrivi in cifre il numero *millecentosette*.

Risposta: ....

(Translation of the authors: Write the number one thousand one hundred and seven in digits. Answer: ,,,,)



Note. In 2009, at grade 5, 'citizenship' is split into two categories (1. native student; 2. Foreign student). In 2016, the same variable has three answer curves (i.e., 1. native student; 2. First-generation student; and, 3. Second-generation student). Therefore, the second graph shows three lines, one per each citizenship status.
Source: our elaboration on INVALSI data

The two items here reported (Table 7) belong to the content domain *Numbers* and both require a switch between two different semiotic registers (Duval, 1993): students have to identify (in the first item) or write (in the second one) the number expressed in words. The amount of text in the questions is limited but the performance of foreign students might be affected by their ability to understand a number stated in words. As already stated, this item meets only one of the criteria (lexical complexity) identified above, but the obstacle caused by the reading of a large number written as a single word might be crucial for foreign students who have not achieved complete mastery of the Italian language. Indeed, numbers (in particular, large numbers with the presence of 0 digits) are difficult words not only for foreign students but also for native students who are linguistically weak. One possible mistake consists in writing the individual numbers one by one, without considering the digits' place values: by following this process, some students might answer the second item with 10001007. Furthermore, in both items the percentage of missing answers is almost null both for foreign students and for natives (suggesting that all the students are confident of understanding and answering the question); the gap highlighted is mostly due to a higher percentage of incorrect answers given by foreign students.

The comparison of DIF plotting of the two questions highlights a similar and interesting trend that confirms the robustness of our analysis. In 2009, at grade 5, for example, the gap between native (blue line) and foreign students (green line) is wider at the lower end of the ability distribution; it narrows when moving from lower to upper levels of the ability trait; and it disappears completely at the top levels. The same item administered seven years later shows exactly the same differences between Italian and foreign students.

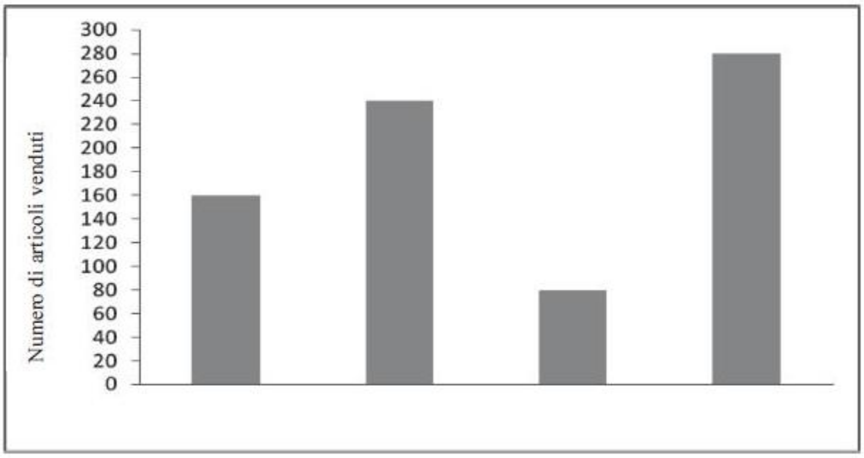
Another interesting comparison of DIF plots of similar items is the one proposed in the following table (Table 8), which includes two grade 6 items.

Table 8

Comparison between item D9 administered at grade 6 in 2010 and item D12 administered at grade 6 in 2012

Item_D9 Grade_6(2010)

Il grafico in figura rappresenta gli articoli venduti da un'edicola nell'ultima settimana, ma i loro nomi sono scomparsi dal grafico. I quotidiani sono stati i più venduti, mentre i CD sono stati i meno venduti; sono stati venduti più settimanali che libri.



Quanti settimanali sono stati venduti?

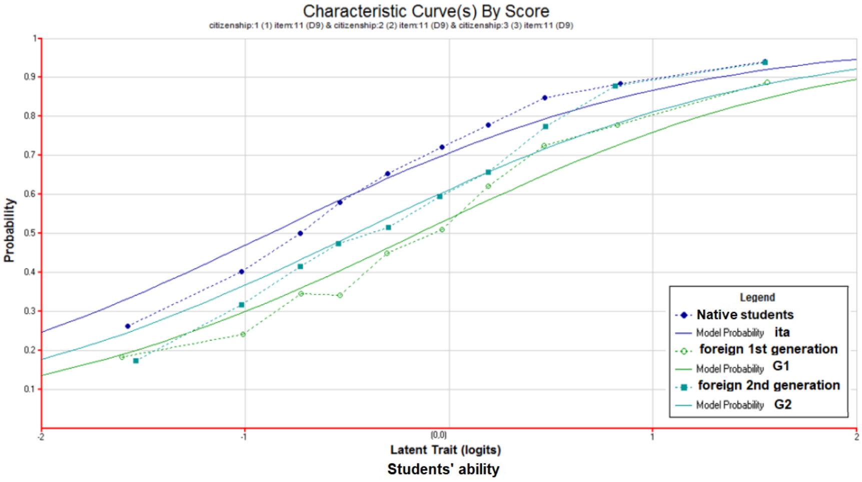
☐ A. 80

☐ B. 160

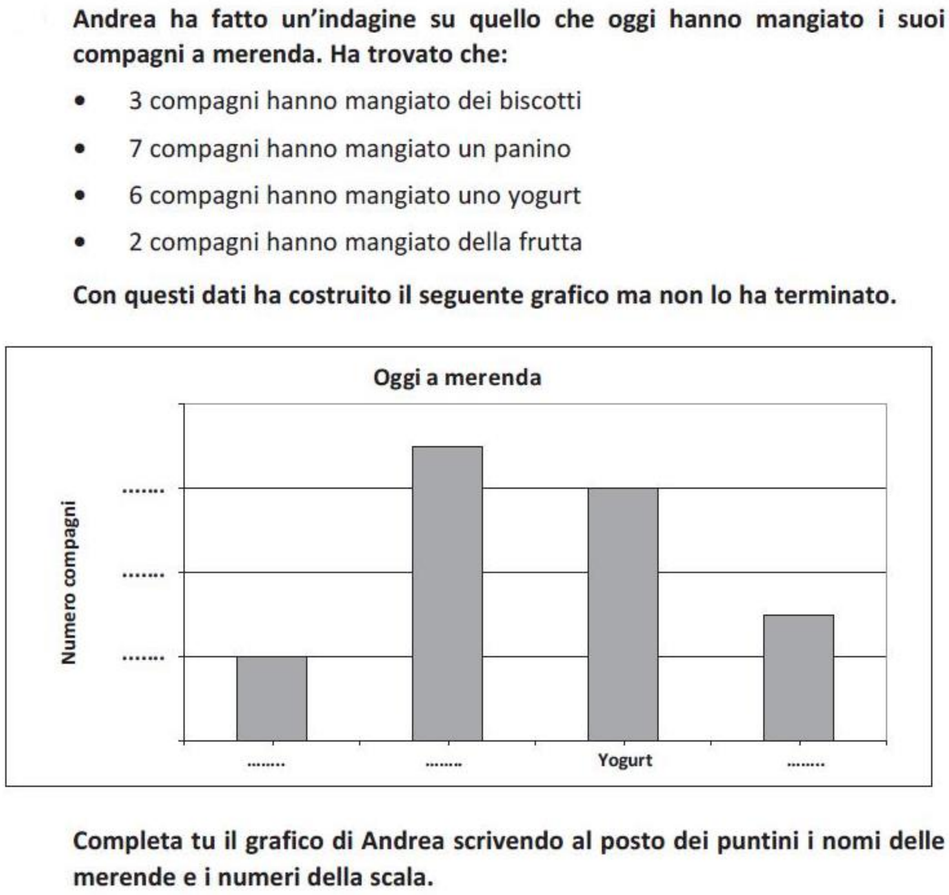
☐ C. 240

☐ D. 280

(Translation: The graph in the figure shows the items sold by a newsagent over the last week, but the item names have disappeared from the graph. Newspapers were the best-selling, while CDs were the least sold; more magazines than books were sold weekly.
How many magazines have been sold?)



Item_D12_Grade_6(2012)



Andrea ha fatto un'indagine su quello che oggi hanno mangiato i suoi compagni a merenda. Ha trovato che:

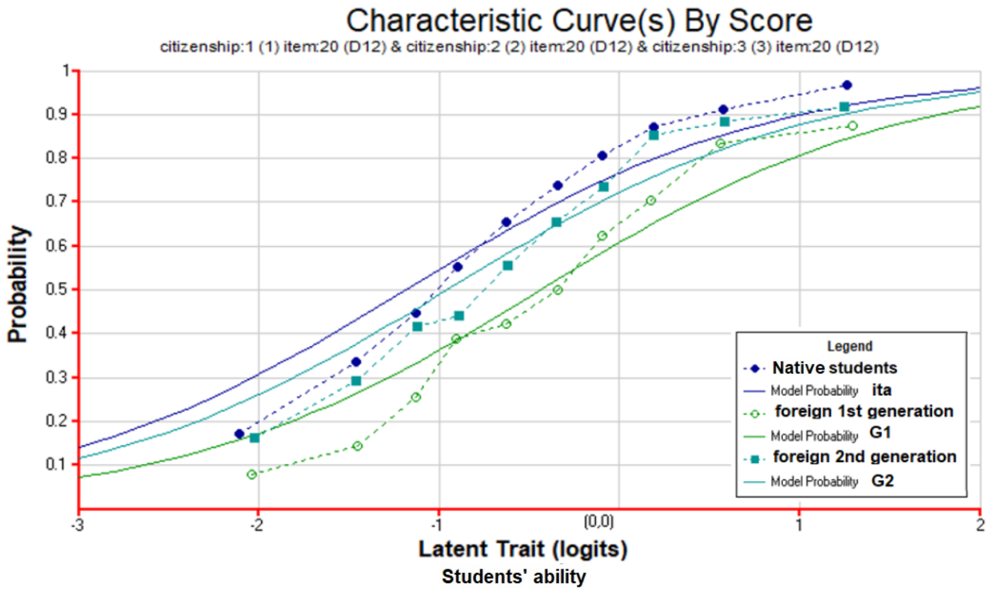
- 3 compagni hanno mangiato dei biscotti
- 7 compagni hanno mangiato un panino
- 6 compagni hanno mangiato uno yogurt
- 2 compagni hanno mangiato della frutta

Con questi dati ha costruito il seguente grafico ma non lo ha terminato.

Oggi a merenda
Numero compagni
....
....
....
....
....
Yogurt
....

Completa tu il grafico di Andrea scrivendo al posto dei puntini i nomi delle merende e i numeri della scala.

(Translation of the authors: Andrea carried out a survey on what his classmates ate for breakfast today. He found that: 3 classmates ate biscuits; 7 classmates ate a sandwich; 6 classmates ate a yogurt; 2 classmates ate fruit; With this data, he built the following graph but did not finish it. Complete Andrea's graph by writing the names of the foods and the numbers of companions on the scale instead of the dots.)



Source: our elaboration on INVALSI data

Both the items belong to the domain *Data and Uncertainty* and students have to complete a graph in which some labels are missing, using the information given in the text. In contrast to the previous questions, in these items the amount of text describing the graph is greater, and it is necessary to read and comprehend the text thoroughly in order to answer correctly. Moreover, the main difficulty for foreign students in these two items might be due to non-mastery of a mixed text which includes a verbal description and a graph in the first item and a verbal description, a list and a graph in the second. While in the first item we have higher syntactic complexity, in the second the list might help foreign students to focus on the individual pieces of data. Finally, in both items we found elements of lexical complexity, such as the use of polysemic terms that are frequently used with other meanings in common language (e.g., ‘settimanali’ -weekly and magazines - and ‘scala’ – scale and stepladder). At grade 6, for item D9 administered in 2010 and item D12 administered in 2012, differences between native and foreign students (both first and second generation) are small at the lowest and highest levels of the ability trait and much bigger in the middle, around 0.00 *logit*, that is for medium-ability students. Also in this case, the comparison of DIF analysis of similar items over time confirms a gap in favour of natives and a similar trend in answering the question for both native and foreign students.

7. Conclusions

Results from our analysis confirmed, as already highlighted in our previous work (Cascella & Giberti, 2020), that the reading demand criteria is not sufficient to explain the difference in performance between Italian and foreign students. Analysis of the citizenship gap in mathematics must also consider other factors that may be compounded and can become an obstacle for students with language weaknesses. Our results confirm that foreign students' difficulties might not be related to the quantity of text so much as to specific obstacles inherent in the text itself in terms of lexicon, syntactic formulation and type of text (e.g., mixed-text which requires a linking of information between the text and graph/figures/etc.). Furthermore, there is also evidence that in some cases there are items in which it is not necessary to read the text, because all the information can be deduced from a figure and the question is easily grasped. More specifically, in the two items reported in Table 7 (D1 grade 5 2009 and D1 grade 5 2016), 'language' is not only a support or mediator of the task, but it is intrinsically part of the task. Students have to decode a word, hence to translate from a verbal register to a symbolic one. This is true also for item D16 (Figure 3) where we have a verbal description of mathematical object. In both cases, language is part of the task, whilst in D24 (Figure 6) or D9 (Figure 4) it is not: language is simply used for the 'storytelling' of the situation. Even heavier is the role of language un D13 (Figure 5), where we have to decode an argumentation and the correctness of the answer depends on a (subtle) decoding of this text. Then there is a qualitative difference in the role of the language in the formulation of the item: there are items where a comprehension of specific words and linguistic structure is needed in order to answer, and items where the used words might be substituted by other, or even avoided without changing the scope of the task.

Starting from the criteria related to the reading demand proposed by Ajello and colleagues (2018), the first step of our study led us to the formulation and validation of six features of items that can be considered when analysing differences between students from native and immigrant backgrounds:

I. large amount of text (more than 100 words)

II. non-continuous text or mixed text (alternation between text, figures, graphs, formulas)

III. necessity of text reading

IV. lexical complexity of the text

V. syntactic complexity of the text

VI. type of the item (we consider that multiple choice questions help foreign students in answering the question while open-ended and argumentative questions might be an obstacle for students with language difficulties)

Finally, we analysed all the items administered by INVALSI, year by year from 2010 onwards, to different students in different schools, years, and regions. We focused our attention on those that display the same characteristics in terms of mathematical content, question intent and formulation. We performed a Differential Item Functioning Analysis to explore possible differences in performance between native and foreign students in order to understand whether the differences we observed in the first step of our research might be attributable only to the particular samples we analysed or indeed if our result can actually be considered ‘sample free’, that is independent from a specific students’ sample.

Results presented in relation to the second research step completely overlap those shown in relation to Step 1, thus confirming the existence of a real relationship between the items here investigated (and thus their mathematical content and their phrasing) and foreign students’ disadvantage. In particular, the comparison between DIF plots based on citizenship highlighted that students’ answering behaviour is similar across years and samples. Results presented in section 3 strengthen those reported in section 2 and highlight the importance of an item-level analysis to identify features and causes of this gap.

We confirm that this work is original and has not been published elsewhere, nor is it currently under consideration for publication elsewhere. We have no conflicts of interest to disclose.

References

Adams, R. J., & Wu, M. L. (2010). *Differential Item Functioning.* Australian Council for Educational Research.

Ajello, A. M., Caponera, E., & Palmerio, L. (2018). Italian students' results in the PISA mathematics test: does reading competence matter? *European Journal of Psychology of Education, 33*(3), 505-520.

Altieri Biagi, M. L. (1978). *Didattica dell'italiano.* Edizioni Scolastiche Bruno Mondadori.

Altieri Biagi, M. L., Frasnedi, F., Pasquini, E., & Speranza, F. (1982). *Un'esperienza interdisciplinare nella scuola media*. Il Mulino.

Altieri Biagi, M. L., Pasquini, E., & Speranza, F. (1979). *Per una didattica interdisciplinare nella scuola media*. Il Mulino.

Altieri Biagi, M. L., & Speranza, F. (1981). *Oggetto, parola, numero. Itinerario didattico per gli insegnanti del primo ciclo*. Nicola Milano.

Andrich, D., Sheridan, B., & Luo, G. (1997-2012). *RUMM2030: Rasch Unidimensional Models for Measurement.* RUMM Laboratory.

Bergqvist, E., Theens, F., & Österholm, M. (2018). The role of linguistic features when reading and solving mathematics tasks in different languages. *The Journal of Mathematical Behavior, 51*, 41-55.

Böhlmark, A. (2008). Age at immigration and school performance: A siblings analysis using Swedish register data. *Labour Economics, 15*(6), 1366-1387.

Bolondi, G., & Cascella, C. (2020). 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. *La matematica e la sua didattica, 28*(2), 255-276.

Bolondi, G., & Ferretti, F. (2021). Quantifying Solid Findings in Mathematics Education: Loss of Meaning for Algebraic Symbols. *International Journal of Innovation in Science and Mathematics Education, 29*(1).

Caponera, E., Sestito, P., & Russo, P. M. (2016). The influence of reading literacy on mathematics and science achievement. *The Journal of Educational Research, 109* (2), 197-204.

Carhill, A., Suárez-Orozco, C., & Páez, M. (2008). Explaining English language proficiency among adolescent immigrant students. *American Educational Research Journal*, *45*(4), 1155-1179.

Cascella, C., & Giberti, C. (2020). Beyond text comprehension: exploring items' characteristics and their effect on foreign students' disadvantage in mathematics, *International Journal of Mathematical Education in Science and Technology, 1-21.* 10.1080/0020739X.2020.1836408

Cascella, C., Giberti, C., & Bolondi, G. (2020). An analysis of Differential Item Functioning on INVALSI tests, designed to explore gender gap in mathematical tasks. *Studies in Educational Evaluation, 64*, 100819.

Chiari, I. (2017). Il vocabolario di base dell'italiano e la società civile. In S. Gentini, G. Solimine, E. Piemontese, & T. De Mauro (Eds.), *Un intellettuale italiano*. Sapienza University Press, pp. 165-172.

Creswell, J. W., & Clark, V. L. P. (2017). *Designing and conducting mixed methods research.* Sage publications.

D'Amore, B. (2000). Lingua, Matematica e Didattica. *La matematica e la sua didattica, 1,* 28- 47.

De Mauro, T. (1994). *Capire le parole.* Laterza.

De Mauro, T. (2016). Il nuovo Vocabolario di Base. *L'internazionale.* Retrieved from https://www.internazionale.it/opinione/tullio-de-mauro/2016/12/23/il-nuovo-vocabolario-dibase-della-lingua-italiana (February 2021).

Duval, R. (1993). Registres de représentations sémiotique et fonctionnement cognitif de la pensée. *Annales de Didactique et de Sciences Cognitives, 5,* 37-65.

Entorf, H., & Lauk, M. (2008). Peer effects, social multipliers and migrants at school: An international comparison. *Journal of Ethnic and Migration Studies, 34*(4), 633-654.

Ferrari, P. L. (2021). *A proposito di Educazione matematica, lingua, linguaggi. Costruire, condividere e comunicare matematica in classe,* Torino, UTET.

Ferretti, F., & Giberti, C. (2021). The Properties of Powers: Didactic Contract and Gender Gap. *International Journal of Science and Mathematics Education, 19*(8), 1717-1735.

Giannelli, G. C., & Rapallini, C. (2016). Immigrant student performance in Mathematics : Does it matter where you come from?. *Economics of Education Review, 52,* 291-304.

Giberti, C., & Viale, M. (2017). Lo studente non madrelingua italiana di fronte al testo delle prove INVALSI di italiano e matematica: dall'analisi dei dati a spunti di intervento. In M. Vedovelli (A cura di), L'italiano dei nuovi italiani - Atti del XIX Convegno Nazionale del

GISCEL. I QUADERNI DEL GISCEL (pp. 343-362). Roma: Aracne Editrice. ISBN: 978-88-255-0034-9

Giberti, C., & Viale, M. (2019). L'impatto del gap linguistico nelle performance degli studenti madrelingua e non madrelingua italiana: dai risultati delle prove INVALSI al lavoro in classe. In P. Falzetti (Ed.), *Uno sguardo sulla scuola. II Seminario "I dati INVALSI: uno strumento per la ricerca".* Milano, Italy: Franco Angeli Editore. ISBN 9788891794796.

INVALSI (2019). *Rapporto Nazionale. Rapporto Prove INVALSI 2019*. INVALSI. Retrieved March 2022 from https://invalsi-areaprove.cineca.it/docs/2019/rapporto_prove_invalsi_2019.pdf

Lavinio, C. (2007). Difficoltà linguistiche in matematica. In R. Imperiale, B. Piochi, & P. Sandri (Eds.), *Matematica e difficoltà: i nodi dei linguaggi*, Pitagora, pp. 15-25.

Morley, L., Leach, F., & Lugg, R. (2009). Democratising higher education in Ghana and Tanzania: Opportunity structures and social inequalities. *International Journal of Educational Development, 29*(1), 56-64.

Mullis, I. V. S., Martin, M. O., & Foy, P. (2013). The impact of reading ability on TIMSS mathematics and science achievement at the fourth grade: an analysis by item reading demands. In M.O. Martin, & I. V. S. Mullis (Eds.), *Relationships among reading, mathematics and science achievement at the fourth grade-implications for early learning.* TIMSS & PIRLS International Study Center.

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). *TIMSS 2019 International Results in Mathematics and Science*. Retrieved from Boston College, TIMSS

& PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/international-results/

OECD (2013). *PISA 2012 Results: Excellence Through Equity: Giving Every Student the Chance to Succeed* (Volume II). OECD Publishing.

OECD (2015). *Can the performance gap between immigrant and non-immigrant students be closed?.* PISA in Focus, No. 53, OECD Publishing, Paris.

OECD (2016). *PISA 2015 Results (Volume I): Excellence and Equity in Education*. OECD Publishing.

Ongaki, N. M., & Musa, F. W. (2014). Enhancing Socio-Economic Equity in Accessing Quality Education: A Case of Form One Selection Policy in KISII County, Kenya. *The International Journal of Business & Management, 2*(11), 157.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning.* Sage Publications.

Radford, L., & Barwell, R. (2016). Language in mathematics education research. In Á. Gutiérrez, G. C. Leder & P. Boero (Eds.), *The Second Handbook of Research on the Psychology of Mathematics Education*, pp. 275–313. Brill Sense.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Danish Institute for Educational Research. (Expanded edition, 1980. University of Chicago Press).

Sbaragli, S., & Demartini, S., (2021). *Italmatica. Lingua e strutture dei testi scolastici di matematica.* Dedalo Editore.

Schleicher, A. (2006). Where immigrant students succeed: a comparative review of performance and engagement in PISA 2003: OECD 2006. *Intercultural Education, 17*(5), 507-516.

Sciumbata, F.C. (2021). Dal plain language all'easy-to-read per lettori con disabilità intellettive: oltre la semplificazione. *Lingue e Linguaggi, 41*, 199-213.

Sutherland, R.J., & Isherwood, T. (2016). The Evidence for Easy-Read for People with Intellectual Disabilities: A Systematic Literature Review. *Journal of Policy and Practice in Intellectual Disabilities*, *13*(4), 297–310.

Viale, M. (2019). *I fondamenti linguistici delle discipline scientifiche. L'italiano per la matematica e le scienze a scuola.* CLEUP Editors.

Wößmann, L., & Schütz, G. (2006). Efficiency and equity in European education and training systems. *Analytical Report for the European Commission prepared by the European Expert Network on Economics of Education, (Bruselas: Comisión Europea, 2006).*

Wright, B. D., Linacre, J. M., Gustafson, J. E., & Martin-Lof, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3).

Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series, 2012*(1), i-30.

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement, 36*(1), 1-28.

Appendix 1

Differential item functioning analysis, results from the previous research (Cascella & Giberti, 2020).

Table I. Differential items' functioning by citizenship status at grade five (2009)

	Native students					**Foreign students**				
Item	**Estimate**	**Error**	**MNSQ**	**CI**	**T**	**Estimate**	**MNSQ**	**CI**	**T**	**DIF magnitude**
D1	-0.133	0.034	0.99	(0,97,1,03)	-0.5	0.133*	1	(0,91,1,09)	0	0.133
D8d	-0.162	0.044	1.00	(0,95,1,05)	0.00	0.162*	1	(0,86,1,14)	0	0.162
D22	0.164	0.023	1.09	(0,99,1,01)	23.5	-0.164*	1.06	(0,97,1,03)	4.6	0.164
D25a	-0.275	0.032	0.98	(0,97,1,03)	-1.4	0.275*	0.97	(0,92,1,08)	-0.8	0.275
D26	-0.139	0.025	0.92	(0,98,1,02)	-10.3	0.139*	0.91	(0,96,1,04)	-4.2	0.139

Separation reliability = 0.906. Chi-square test of parameter equality = 380.29, df = 40, Sig. level = 0.000
Source: Our elaboration based on the INVALSI data.

Table II. Differential item functioning by citizenship, at grade six (2010)

	Native					First-Generation (G1)					Second-Generation (G2)					DIF magnitude		
Item	Estimate	Error	MNSQ	CI	T	Estimate	Error	MNSQ	CI	T	Estimate	Error	MNSQ	CI	T	ITA v. G1	ITA vs. G2	G1 vs. G2
D6	-0.305	0.08	0.98	(0.98,1.02)	-1.6	0.277	0.08	1.06	(0.83,1.17)	0.8	0.028	0.08	1.05	(0.83,1.17)	0.6	-0.582	-0.333	0.249
D8_a	-0.194	0.03	0.97	(0.99,1.01)	-5.9	0.248	0.03	0.93	(0.97,1.03)	-5.1	-0.054	0.03	0.96	(0.96,1.04)	-1.9	-0.442	-0.140	0.302
D8_b	-0.117	0.03	0.97	(0.99,1.01)	-7.5	0.175	0.03	0.96	(0.96,1.04)	-2.2	-0.058	0.03	0.98	(0.96,1.04)	-1.3	-0.292	-0.059	0.233
D9	-0.151	0.02	0.93	(0.99,1.01)	-14.5	0.125	0.03	0.94	(0.97,1.03)	-4.7	0.027	0.03	0.93	(0.96,1.04)	-3.6	-0.276	-0.178	0.098
D10_d	0.253	0.02	1.22	(0.99,1.01)	52.6	-0.261	0.03	1.22	(0.97,1.03)	15.2	0.008	0.03	1.18	(0.96,1.04)	9.5	0.514	0.245	-0.269
D11_b	0.171	0.02	1.13	(0.99,1.01)	31.2	-0.221	0.03	1.11	(0.97,1.03)	7	0.050	0.03	1.13	(0.96,1.04)	6.5	0.392	0.121	-0.271
D13	-0.140	0.02	0.95	(0.99,1.01)	-12.5	0.136	0.03	0.99	(0.97,1.03)	-0.6	0.004	0.03	0.98	(0.96,1.04)	-0.8	-0.276	-0.144	0.132
D14	0.137	0.02	0.96	(0.99,1.01)	-11.1	-0.123	0.03	0.96	(0.97,1.03)	-3.1	-0.014	0.03	0.93	(0.96,1.04)	-4	0.260	0.151	-0.109
D16	-0.132	0.03	0.96	(0.99,1.01)	-7.5	0.139	0.03	0.95	(0.97,1.03)	-3.6	-0.007	0.03	0.97	(0.96,1.04)	-1.3	-0.271	-0.125	0.146
D19	0.171	0.03	1.02	(0.99,1.01)	3.4	-0.176	0.04	1.05	(0.94,1.06)	1.7	0.004	0.03	1.04	(0.93,1.07)	1.1	0.347	0.167	-0.180
D22	0.147	0.03	1.04	(0.99,1.01)	8.9	-0.118	0.03	1.02	(0.97,1.03)	1	-0.030	0.03	1.03	(0.96,1.04)	1.4	0.265	0.177	-0.088
D27_a	-0.035	0.03	0.98	(0.99,1.01)	-2.5	0.143	0.04	0.96	(0.97,1.03)	-2.5	-0.109	0.03	1.03	(0.94,1.06)	1.1	-0.178	0.074	0.252
D27_b	-0.253	0.03	0.93	(0.99,1.01)	-10.8	0.265	0.03	0.92	(0.97,1.03)	-5.7	-0.012	0.03	0.95	(0.95,1.05)	-2	-0.518	-0.241	0.277

Separation Reliability = 0.921. Chi-square test of parameter equality = 1147.14, df = 82, Sig. level = 0.000. ^Quick standard errors have been used. DIF magnitudes greater than 0.43 are highlighted in gray cells (Zwick, 2012).
Source: Our elaboration based on the INVALSI data.

Table III. Differential item functioning by citizenship, at grade eight (2012)

Native	First-generation (G1)	Second-generation (G2)	DIF magnitude

Item	Estimate	Error^	MNSQ	CI	T	Estimate	Error^	MNSQ	CI	T	Estimate	Error^	MNSQ	CI	T	ITA vs. G1	ITA vs. G2	G1 vs. G2
D1	-0,108	0,015	1	(0,91;1,09)	0,1	0,249	0,042	1,02	(0,79;1,21)	0,2	-0,141	0,044	0,98	(0,58;1,42)	0	-0,4	0,0	0,4
D2_b	0,124	0,013	1,04	(0,96;1,04)	2	-0,144	0,038	1,03	(0,88;1,12)	0,5	0,02	0,04	1,01	(0,82;1,18)	0,2	0,3	0,1	-0,2
D3_b	-0,011	0,011	1,05	(0,99;1,01)	9,3	0,14	0,031	1,04	(0,97;1,03)	2,1	-0,129	0,033	1,07	(0,95;1,05)	2,3	-0,2	0,1	0,3
D4_b	-0,078	0,011	1,02	(0,99;1,01)	3,2	0,191	0,032	1,02	(0,97;1,03)	1,2	-0,113	0,033	0,98	(0,93;1,07)	-0,7	-0,3	0,0	0,3
D6	0,245	0,011	1,08	(0,98;1,02)	9,8	-0,312	0,033	1,15	(0,94;1,06)	4,7	0,067	0,035	1,12	(0,91;1,09)	2,5	0,6	0,2	-0,4
D9_a	-0,141	0,011	0,95	(0,99;1,01)	-9,6	0,134	0,033	0,93	(0,94;1,06)	-2,1	0,006	0,035	0,93	(0,93;1,07)	-1,8	-0,3	-0,1	0,1
D10_b	0,16	0,011	1,07	(0,99;1,01)	10,3	-0,024	0,033	1,07	(0,94;1,06)	2,4	-0,136	0,035	1,11	(0,93;1,07)	2,9	0,2	0,3	0,1
D11	-0,05	0,011	0,94	(0,98;1,02)	-7,3	-0,117	0,035	0,97	(0,93;1,07)	-0,8	0,166	0,036	0,96	(0,89;1,11)	-0,7	0,1	-0,2	-0,3
D19_b	0,134	0,01	1,06	(0,99;1,01)	13,3	-0,166	0,031	1,08	(0,97;1,03)	4,4	0,033	0,033	1,07	(0,95;1,05)	2,8	0,3	0,1	-0,2
D24	-0,171	0,01	0,91	(0,99;1,01)	-19,1	0,187	0,031	0,88	(0,96;1,04)	-6,6	-0,016	0,033	0,92	(0,95;1,05)	-3,5	-0,4	-0,2	0,2

Separation Reliability = 0.945. Chi-square test of parameter equality = 3834.43, df = 90, Sig. level = 0.000. ^Quick standard errors have been used. DIF magnitudes greater than 0.43 are highlighted in gray cells (Zwick, 2012).

Source: Our elaboration based on the INVALSI data