Interpreting the gap between foreign and native speaker students in national standardised assessment

Marta Barbero, Clelia Cascella, Chiara Giberti, Matteo Viale

Interpreting the gap between foreign and native speaker students in national standardised assessment

Marta Barbero[1], Clelia Cascella[2], Chiara Giberti[3] and Matteo Viale[4]

[1]University of Applied Sciences and Arts of Southern Switzerland (SUPSI)

[2]University of Manchester, United Kingdom; INVALSI

[3]University of Bergamo, Italy; chiara.giberti@unibg.it

[4]University of Bologna, Italy

The paper is based on the interweaving of qualitative and quantitative analyses to evaluate the learning gap between foreign and native speaker students in the Italian national standardised assessments. The study shows that a linguistic and structural analysis of the items based on specific criteria allows for an initial justification of the resulting differences identified between the students' sub-groups, partially explaining the disadvantages of foreign students.

Keywords: Standardised assessment, citizenship gap in mathematics, differential item functioning.

Introduction

Several research over the last decades have highlighted how students' performances in mathematics is influenced by different factors such as their social and economic status, their gender, their mother tongue or other factors related to the educational system, the characteristics of schools or the structure of curricula (Ongaki & Musa, 2014). Recent research has emphasised several elements conditioning the gap in academic performance among foreign students compared to native speakers: the social and economic situation of families, and the level of education, occupation, or parental expectations (e.g. Schnepf, 2007), the language spoken at home and the resulting language barriers that emerge in the school context (Carhill et al., 2008), the age of arrival in the destination country, or being second-generation immigrants (Schleicher, 2006). These factors play an important role with regard to the knowledge, or lack thereof, of the educational system and its characteristics, hindering the integration of students and the possibility of making full use of their potential (Wößmann & Schütz, 2006). Finally, it is interesting to analyse the language barriers factor: foreign students have to learn mathematics in a language different from their mother tongue while also having to adapt to different cultural practices (Phakeng, 2016). Linguistic diversity must be considered not only as a vocabulary problem, but also as a different way of constructing meanings and new concepts; for immigrant students perceptions and interpretations of mathematics differ from those of native students, influencing the way they experience what happens in the classroom (Gorgorió & Planas, 2001).

Theoretical framework: Citizenship gap in standardised assessment

The results of international surveys such as PISA and TIMSS and recent research (Ajello et al., 2018; Sbaragli & Demartini, 2021) have confirmed the strong link between reading skills and performance in mathematics. A systematic analysis of the language difficulties encountered by students when dealing with mathematical texts suggests that mathematical competences are closely interconnected with reading and text comprehension competences (Sbaragli & Demartini, 2021).

Considering the Italian language, texts used in mathematics (both in standardised tests and in everyday practice) have several characteristics that contribute to their complexity: they are non-continuous or mixed, with an interweaving of texts, images, formulae, etc. which requires a specific cognitive effort; a syntax with a preference for coordination normally prevails, but there is a great deal of implicit subordination, especially in the gerund form; the expression of actions with nouns instead of verbs (nominalisation) is often used; the used language is not always familiar to students by resorting to a progressively increasing number of technical terms that are interconnected; terms from the common lexicon are often used to refer technically to mathematical objects (e.g. 'edge', 'angle'), giving rise to misconceptions and constant difficulties (Sbaragli & Demartini, 2021). Thus, foreign students' disadvantages[1] in mathematics are certainly related to language difficulties but these difficulties are due to a variety of factors as highlighted in previous research (Giberti & Viale, 2019). Considering data from Italian standardised assessment, which are administered every year in different grades from primary to upper secondary schools, in our previous works we performed a pseudo-longitudinal study (i.e. a repeated cross-sectional study of the same birth cohort; see Steel 2011 for an application) to deeper understand the reasons of the citizenship gap in mathematics standardised assessments (Cascella & Giberti, 2020; Cascella et al., 2022). In Cascella & Giberti (2020) we showed that a relationship between text comprehension (as defined by Ajello et al., 2018) and mathematical ability exists and could explain part of the citizenship gap. Nonetheless, some items show a statistically significant Differential Item Functioning (DIF) by citizenship status in favour of native students regardless of their medium or low reading demand level (Cascella & Giberti, 2020). We thus concluded that criteria identifying reading demand are not exhaustive to explain the citizenship gap.

To go a step further, in Cascella et al. (2022), we performed a mixed-method analysis of items showing a statistically significant DIF by citizenship in favour of native students, regardless of items' reading demand, and proposed six new criteria to "classify" texts, as listed here below:

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 (is it possible to answer correctly considering only figures and numbers reported in the item?);
IV. lexical complexity of the text (e.g. presence of terms which do not belong to the 'basic vocabulary' of Italian language (Chiari, 2017) or polysemic terms);
V. syntactic complexity of the text (e.g. use of passive forms, gerund subordinates, etc.);
VI. type of item (we consider that multiple choice items help foreign students while open-ended and argumentative questions might be an obstacle for students with language difficulties).

These criteria are our proposed adaptation of the reading demand criteria (Ajello et al., 2018). In some cases, we modified the threshold of the criteria (e.g. number of words which was originally 200), other criteria were newly added (e.g. crit. III). These modifications allowed to explain the DIF by citizenship. Results based on different items with similar characteristics showed DIF by citizenship, thus supporting our interpretative hypotheses (Cascella et al., 2022).

[1] In this and in previous works, 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).

Aim and methods

Building on our previous research and on the criteria already proposed (Cascella et al., 2022), the current was aimed at validating the criteria on a new subset of data to figure out if these criteria are useful to interpret learning gaps between foreign and native speaker students. Indeed, in previous research the criteria were formulated on the basis of items showing DIF in favour of native students, while in the current paper we analyse items from italian standardised assessments by considering DIF and our criteria in parallel, within the framework of a blind analytical strategy: part of the research team (1 researcher) analysed data from a quantitative point of via the Differential Item Functioning analysis, another part of our research team (2 researchers), applying our qualitative criteria (Cascella et al., 2022) to all items of the test. The fourth member of our team served as "external" judge. Then, results from the quantitative and qualitative analysis were compared in order to answer the following question: Do the criteria proposed by Cascella and colleagues (2022) explain the citizenship gap also performing a parallel quantitative-qualitative analysis of a new subset of data?

While, in our previous study we performed a quali-quantitative analysis to explain students' gap in mathematics by citizenship, the current paper aims to validate the criteria proposed in Cascella et al. (2022) by using those criteria to analyse a different set of data. To validate our proposed interpretative criteria we have to understand if the items highlighting citizenship DIF in a new set of data are characterised by the same features and, to this end, we performed the two analyses separately.

Data

For the purposes of the present study, we analysed data collected by the Italian national institute for the evaluation of the educational system (hereafter INVALSI), at grade 5 in 2017. The achievement test administered in 2017 consisted of 39 (open-ended or multiple-choice) items spreading into a wide range of difficulty. The INVALSI sample was composed of 25,482 students, approximately the 2% of these students are first generation foreign students and the 7% are second generation students.

The methodological strategy

In the present study, we employed a quali-quantitative strategy aimed at comparing results from a quantitative analysis (step 1) with the qualitative criteria presented in the previous paragraph (step 2). Step 1 and 2 were run simultaneously and independently of one another to avoid reciprocal influences.

Step 1 (quantitative analysis). Data were analysed by using the Rasch model (1960/1980) that estimates the probability of encountering each item successfully as a function of students' ability compared with item's difficulty. Within the framework of the Rasch model, it is assumed that students' personal characteristics do not play any role in affecting such a probability: therefore, if the model holds from a statistical point of view (i.e., if appropriate levels of data-model fit are verified), then students' estimated ability is invariant across sub-groups of people (e.g., males versus females, native versus foreign students, and so on). Violations of measurement invariance can be detected by performing a Differential Item Functioning (DIF) analysis. DIF can be due to weaknesses in the instrument used to measure students' ability or it can represent genuine differences between sub-groups of students. In regards to the latter case, previous studies (Cascella et al., 2020) argued that, when DIF is not detrimental for measurement, it can be used as a tool to identify didactical

interpretation of learning gaps between sub-groups of students. In fact, it is worth noting that even if statistically significant, DIF can be negligible in SIZE (or magnitude), as summarised by Zwick (2012): lower than 0.43 = negligible; between 0.44 and 1 = moderate; and greater than 1 = high.

Step 2 (qualitative analysis). Each item of the test was categorised according to the 6 criteria listed in the previous section. Then the quantitative evaluation was compared with the qualitative results. A case study of the items identified as those playing the largest role in explaining the citizenship gap was carried out to accomplish the analysis.

Results

Quantitative results

The DIF analysis highlighted that most of the items showed a statistically significant DIF. Yet, this DIF is not detrimental for measurement as all items' DIF magnitude were far below the 0.43 threshold (e.g., Zwick, 2012), thus showing a negligible DIF. Table 1 reports the items showing a statistically significant DIF.

Qualitative results

The clarity and unambiguity of the criteria proposed to classify items (Cascella et al., 2022) is confirmed by the very low rate (3%) of disagreement between the researchers. The 7 cases in which the two researchers involved in the qualitative analysis didn't agree were analysed and defined by the external judge. The classification of all the items of the test highlights that only two items are considered with a large amount of text (Crit. I: D2, D19) while almost all items except 6 (D4, D5, D12, D17, D23, D26) have mixed text (Crit. II). All items include some text but in 4 items this text was not necessary (Crit. III) to answer the question (D9, D16, D29, D30). In almost half of the cases, we identify elements of lexical complexity (Crit. IV) such as words which do not belong to the 'basic vocabulary' (Chiari, 2017) and polysemic words. A syntactic complexity (Crit. V) was found only in 7 items (D5, D11, D19, D20, D22a, D22b, D26) and 16 items out of the 39 were multiple choice items while the other required an open answer, in few cases argumentative (Crit. IV).

Cross-over results

In this paper we focus on the 16 out of 39 items (Table 1) highlighting a slightly significant DIF between Italian students and 1st generation foreign students to understand if the proposed criteria are useful to explain citizenship disadvantages. Within these items, 6 show also a DIF between Italian students and 2nd generation foreign students, and just 1 highlights a DIF only between Italian and 2nd generation foreign students (Table 1). Almost all these items meet at least 3 out of the 6 criteria proposed. Item D9 is the only one meeting 2 criteria and no one meets less than 2 criteria. Item D9 is a particularly difficult item, almost all the students struggle in answering it and the fact that, usually, only few 1st generation foreign students are top performers suggests that the high difficulty could be the reason for DIF. It is interesting to observe that all the items identified and reported in Table 1 are open answers items, except item D2, a multiple choice item in which criteria I, II, and III could explain citizenship gap. As previously mentioned, almost all INVALSI items include less than 100 words; in grade 5 2017, test here analysed, the only 2 items with a large amount of texts are item D2 and D19, the former is the one already mentioned as an item that create a gap due to several reasons

while the latter does not show a significant DIF. Approximately half of the selected items evidence lexicon difficulties and the syntactic complexity appears only in 4 items.

Table 1: Analysis of the item with significant DIF between Italian and 1st generation foreign students [2]

ITEM	I. large amount of text	II. non-continuous text	III. necessity of text reading	IV. lexical complexity	V. syntactic complexity	VI. type of item[3]	DIF ITA-2gen	DIF ITA-1gen
D2	yes	yes	yes	no	no	MC		x
D4	no	no	yes	no	yes	OA	x	x
D5	no	no	yes	no	yes	OA		x
D7	no	yes	yes	yes	no	OA		x
D9	no	yes	no	no	no	OA		x
D10a	no	yes	yes	no	no	OA		x
D10b	no	yes	yes	no	no	OA	x	
D13a	no	yes	yes	yes	no	OA		x
D13b	no	yes	yes	yes	no	OA		x
D15a	no	yes	yes	no	no	OA	x	x
D15b	no	yes	yes	no	no	OA	x	x
D15c	no	yes	yes	no	no	OAj		x
D21	no	yes	yes	yes	no	OA	x	x
D22a	no	yes	yes	yes	yes	OA	x	x
D22b	no	yes	yes	yes	yes	OA	x	x
D33	no	yes	yes	yes	no	OA	x	x

Discussion

The current paper builds on our previous research aimed at exploring the citizenship gap in mathematics and thus identifying criteria that could explain the citizenship gap. In this research we validate these criteria through a parallel quantitative-qualitative analysis on a new subset of data. We found that just two items showing DIF were characterised by a large amount of text. Such a result could suggest to further modify (at least some of) the criteria proposed by Cascella and colleagues (2022): for instance, the criteria I (number of words) could be based on more levels (e.g. low, medium and high number of words) rather than a single threshold. Instead, criteria II and III are confirmed as significant to explain the foreign students' disadvantage: almost in all the items identified through DIF there is the necessity to read and understand the text and coordinate the information gathered by the text with other representations (figures, tables, etc.).

Our results confirm the influence of the items' format (Crit. VI) on foreign students' answers: almost all items with a significant DIF are open answers. Open answer questions thus represent an element

[2] An 'x' in the last two columns indicates that the item was perceived as more difficult by 'second-generation' (penultimate column) or by 'first-generation' students (in the ultimate column), compared with native students. In the present study we did not highlight items showing a statistically significant DIF favouring foreign against native students.

[3] MC = multiple choice; OA = open answer, OAj = open answer requiring justification

of difficulty not associated with (foreign) students' ability in Mathematics, and that is thus captured by the Rasch model as a DIF. Furthermore, the significant DIF highlighted in relation to 1st generation foreign students in the only multiple choice item identified (D2, Figure 1), could be deeply explained following the other criteria: D2 is one of the two items with a large amount of text, it is a non-continuous text (the information in the graph need to be connected to the text which explain the context and to the sentences reported in the true/false table) and there is the necessity of reading it.

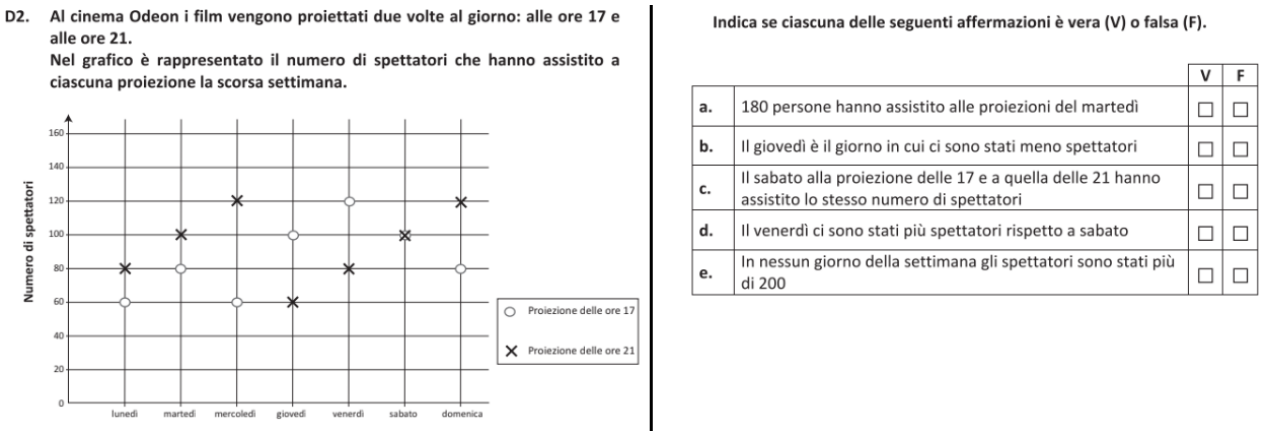

Figure 1: text of item D2[4]

Finally, as a result we observe that the DIF emerged in the identified items could be deeply explained by the criteria proposed. We report here, as a paradigmatic example, item D21 (Figure 2a) which showed a significant DIF that could be explained by several factors, that are (i) the necessity of reading the text and connecting the information gathered in it and in the figure, (ii) the polysemic word '*scala'* ('scale' which in Italian also means 'stair'), and (iii) the open answer request.

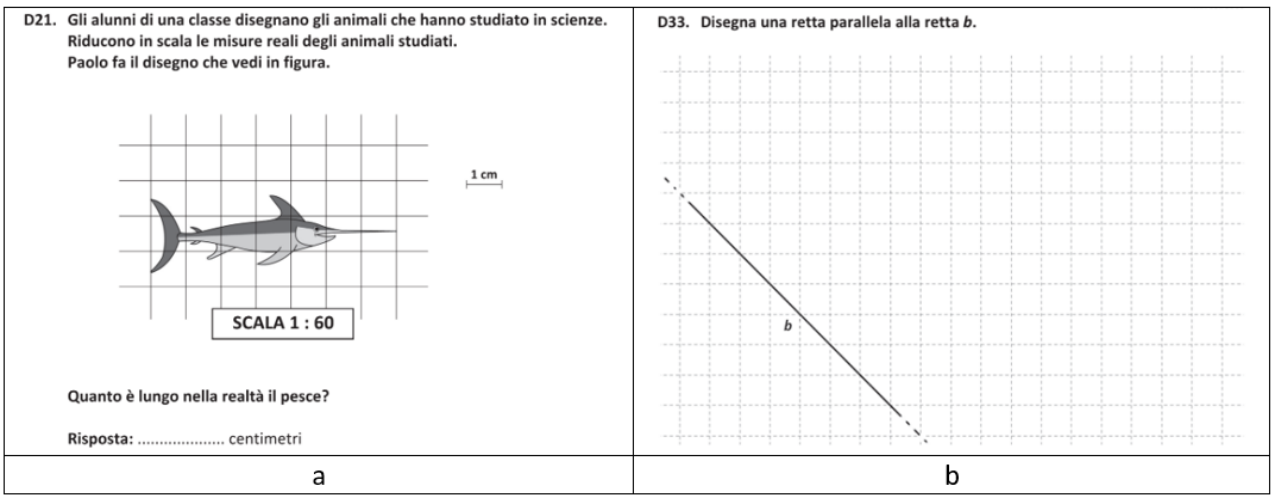

Figure 2: text of the D21 and D33 items[5]

[4] Stimulus translated by authors: "At the cinema the films are screened twice a day: at 17 and 21. The graph shows the number of spectators who watched each screening last week. Decide if each of the following statements is true or false."

[5] Stimulus translated by authors: D21 text - "Pupils draw the animals they studied in science. They scale down the actual measurements of the animals they studied. Paolo makes the drawing you see in the picture. How long is the fish in reality?"; D33 text - "Draw a line parallel to line b")

Another interesting example coming from our analysis is item D33 (Figure 2b). This item shows a statistically significant DIF both for 1st and 2nd generation foreign students, despite a very bare text. Nonetheless, according to our criteria, it is worth noting that this is an open answer question in which the text reading is necessary, non-continuous, and has words from a specific vocabulary ('*retta'* which means '*line'* and '*parallela'* which means '*parallel'*).

Furthermore, we also observed that we can use these criteria to deepen our understanding of items highlighting a citizenship gap, but we cannot state that items following these criteria, are, as a consequence, items creating a citizenship gap. Indeed, we observe 5 items (D20, D10c, D19, D28, D11) following four or more of these criteria but without evidencing a significant DIF.

Conclusions

In developing an achievement test, one of the challenges to face is constructing invariant items by students' characteristics (such as their citizenship status). Nonetheless, especially in Large-Scale Assessment, observing statistically significant DIF is quite common. DIF is not always a cause of concern. For example, it can be not detrimental for measurement when it is below some thresholds commonly used in educational literature (e.g., Zwick, 2012). Results presented showed that, in INVALSI data, there was no detrimental DIF, thus guaranteeing the goodness of measurement. Furthermore, a gap with a significant DIF can be explained by the linguistic and structural analysis of the item: the structure of the item with a non-continuous text, the necessity of text reading, the lexical complexity and the open answer item format are the criteria that most explain the foreign students' gap. These results prompt us to develop a further in-depth analysis to study the factors that influence this gap to understand the variables that determine the disadvantages of foreign students.

References

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. https://doi.org/10.1007/s10212-018-0385-x

Carhill, A., Suárez-Orozco, C., & Páez, M. (2008). Explaining English language proficiency among adolescent immigrant students. *American Educational Research Journal, 45*(4), 1155–1179. https://doi.org/10.3102/0002831208321443

Cascella, C. (2020). Intersectional effects of socioeconomic status, phase and gender on mathematics achievement. *Educational Studies, 46*(4), 476–496. https://doi.org/10.1080/03055698.2019.1614432

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, 53*(6), 1588–1608. http://dx.doi.org/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. https://doi.org/10.1016/j.stueduc.2019.100819

Cascella, C., Giberti, C., & Viale, M. (2022). Investigating foreign students' disadvantage in mathematics: A mixed method analysis to identify features of items favouring native students. *The Journal of Mathematical Behavior, 67*, 100990. https://doi.org/10.1016/j.jmathb.2022.100990

Chiari, I. (2017). Il vocabolario di base dell'italiano e la società civile [The basic vocabulary of Italian and society]. In S. Gentini, G. Solimine, E. Piemontese, & T. De Mauro (Eds.), *Un intellettuale italiano* [An Italian intellectual] (pp. 165–172). Sapienza Univ. Press.

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 [The impact of the language gap in the performance of native and non-native Italian speakers: from INVALSI test results to classroom work]. In P. Falzetti (Ed.), *Uno sguardo sulla scuola. II Seminario "I dati INVALSI: uno strumento per la ricerca"* [A look at schools. The seminar 'INVALSI data: A tool for research'] (pp. 117–140). Franco Angeli Editore.

Gorgorió, N., & Planas, N. (2001). Teaching mathematics in multilingual classrooms. *Educational Studies in Mathematics, 47*(1), 7–33. https://doi.org/10.1023/A:1017980828943

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. *The International Journal of Business & Management, 2*(11), 157.

Phakeng, M. S. (2016). Mathematics education and language diversity. In A. Halai & P. Clarkson (Eds.), *Teaching and learning mathematics in multilingual classrooms. Issues for policy, practice and teacher education* (pp. 11–23). Sense Publishers. https://doi.org/10.1007/978-94-6300-229-5_2

Rasch, G. (1960/1980). *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* [Italmatica. Language and structures of mathematics textbooks]. Dedalo Editore.

Schleicher, A. (2006). Where immigrant students succeed: A comparative review of performance and engagement in PISA 2003. In OECD (Eds.) *Intercultural education, 17* (pp. 507–516). https://doi.org/10.1787/9789264023611-en

Schnepf, S. V. (2007). Immigrants' educational disadvantage: An examination across ten countries and three surveys. *Journal of Population Economics, 20*(3), 527–545. https://doi.org/10.1007/s00148-006-0102-y

Steel, D. (2011). Repeated cross-sectional design. In P. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 715–716). Sage.

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*. Comisión Europea.

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, 1*, i–30. https://doi.org/10.1002/j.2333-8504.2012.tb02290.x