Investigating gender differences in mathematics by performance levels in the Italian school system

**Investigating Gender Differences in Mathematics by Performance Levels in the Italian School System**


## Abstract

Amongst the extensive research conducted about girls' lack of interest in STEM subjects, an issue that is frequently examined is the learning of mathematics. This research investigates the gender gap in mathematics among Italian students based on performance levels, using standardised large-scale data from INVALSI tests. We performed a quantile regression to better understand the differences along the entire score distribution. A latent class model was then estimated to identify groups of students with similar performance levels, taking into account the gender covariate. The results indicate that boys are already ahead from primary education and that there is a general decrease in the performance as students progress through the education stages.



**Keywords**: Gender gap, mathematics performance, INVALSI standardised tests, quantile regression, latent class analysis.

## 1. Introduction

The issue of gender gap has long been a question of great interest in a wide range of fields. Within the educational system at international level, the debate has long been focused on the lack of interest in science, technology, engineering and mathematics (STEM) subjects among girls, who tend to prefer the humanities and, as a result, often end up working in the education, healthcare and social services sectors (OECD 2017a; 2019a; UNESCO, 2017). In Italy, where the gender index is deemed acceptable, the situation is hardly ideal. In fact, despite females represent the 58.7% of the total number of graduates in 2019 (according to AlmaLaurea, 2020, the total number of graduates in Italy in 2019 is equal to 290,224), the proportion of female students into the STEM educational fields is still very low. In particular, within bachelor degrees, females represent only the 26.4% of students in engineering and the 26.7% of students in the scientific area[1] (AlmaLaurea, 2020). On the other hand, females are about the 93.8% of students in the educational field (AlmaLaurea, 2020). These gender disparities are a matter of considerable concern as they have a relevant impact on job opportunities, also considering that careers in the STEM field are among the highest-paying.

Amongst the extensive research conducted in this area, an issue that is frequently examined is the role of pre-university education and of the study of STEM subjects, in particular mathematics. The impact of how this discipline is taught, along with the relative attitudes to and interest in this subject, seem to be the main factors pushing young people, and girls in particular, away from the study of STEM subjects (Peterson & Fennema, 1985; Robinson-Cimpian et al., 2014; Sansone, 2014). In the literature, evidences about when the gender gap is revealed first during the scholastic path has been subject to considerable discussion (see, among others, Bedard & Cho, 2010; Fennema, 1974; Hyde et al., 1990; Meinck & Brese, 2019; Muzzatti & Agnoli, 2007; Peterson &

[1] In the Italian university system, the 'scientific area' includes astronomy, physics, mathematics, materials science and computer science.

Fennema, 1985). Moreover, various perspectives (for a review, see Erdoğan et al., 2011; Halpern et al., 2007) have been adopted for attempting to explain the reasons of gender-related disparities in performance ranging from theories based on biological fundamentals (Geary, 1996) to a more cultural approach influenced by gender stereotypes (Nguyen, & Ryan, 2008) or by the social and cultural context (Contini et al., 2017). Also, some studies suggest that the disparities in gender may be related to differences in home and school socialisation of boys and girls (Hadjar et al., 2014).

Many studies conducted on gender differences in mathematics make use of data from large-scale standardised tests and highlight the existence of a significant gap in most countries. Examples are, at international level, the Organization for Economic Co-operation and Development (OECD) Program for International Student Assessment (PISA) or the International Association for the Evaluation of Educational Achievement (IEA) Trends in International Mathematics and Science Study (TIMSS) (see, e.g., Baye & Monseur, 2016; Bedard & Cho, 2010; De Simone, 2013; Karakolidis et al., 2016; Meinck & Brese, 2019; Neuschmidt et al., 2008). At the Italian national level, we find large-scale assessments conducted by the National Institute for the Evaluation of the Education and Training System (INVALSI) (see, e.g., Cascella et al., 2020; Contini et al., 2017; Costanzo & Desimoni, 2017). In the literature, most studies focused on comparing the differences in mean performances (see, e.g., Bedard & Cho, 2010; De Simone, 2013), while only a few studies explored the differences in lower and higher levels of performance in depth (Baye & Monseur, 2016; Contini et al., 2017; Costanzo & Desimoni, 2017; Halpern et al., 2007; Meinck & Brese, 2019; Strand et al., 2006).

The main purpose of this paper is to investigate the learning results in mathematics of young Italian students focusing particularly on the differences in top and low performances. The analysis is based on the INVALSI assessments, which collect response data and collateral information on Italian students for different school grades.

The research questions of this paper concern the following hypotheses.

1. Are there significant differences between boys and girls in their performance in mathematics both on a mean level and at different performance levels? In particular, may the difference in the performance between males and females change for different achievement levels? To verify this hypothesis, two statistical techniques, one based on the regression model and one based on a classification procedure, are applied.
2. Does the gender gap show a similar behaviour within the different school grades? To answer this question, the analysis is conducted on the data for grades 5, 8 and 10.
3. Are there contents in which the highest and lowest levels of performance can be differentiated by gender? To answer this question, the results of the previous steps are also explored by taking into account the mathematics domains.

To verify all the research questions in this study, an in-depth analysis through a multiway statistical approach is conducted allowing to depict multifaceted learning elements. Firstly, for a comparison with the literature results, we evaluated the mean difference of the estimated ability by INVALSI for females and males. Successively, using a quantile regression approach, we evaluated the gender differences at specified levels of the ability distribution (e.g., the median, the $10^{th}$ quantile, the $75^{th}$ quantile). The hypothesis to verify is that the differences are more evident for the lower or higher achievement levels, in comparison to the medium performance level. The results are useful to discover possible different patterns in the tails of the distribution. Finally, a latent class analysis was performed on the test to identify groups of students with similar performance behaviour. This analysis may confirm or not whether the gender differences are more evident within different performance groups. Specifically, three groups were identified (Low, Medium and High) and the gender impact on these groups was analysed. This approach offers a focused solution to translate certain benchmarks of proficiency levels.

Furthermore, while the data has not yet been tracked at a sufficient number of points in time to produce any meaningful longitudinal data, this paper investigates whether it is possible to identify typical characteristics in the first school cycle and how these evolve as students progress through the different educational cycles. The hypothesis is to verify whether the differences between boys and girls are significant and on which aspects said differences are mainly focused, both at the school grade and with regard to subject content.

First of all the paper contributes to confirm, as far as Italian students are concerned, the existence of overall mean gender differences for all grades, as well-known in the national and international literature. In addition, the use of a combination of several statistical methods allows to highlight differences in gap for both different grades and different levels of performance. At grade 5, at the end of primary school, the gap is already established, it is significant from low-medium performance levels remaining significant with a stable impact for the following ability levels. On the other hand, at grade 8, the gender gap is significant already for very low ability levels, showing successively a stable impact from the medium levels. At grade 10, the difference in mathematics performance by gender is significant from the mean ability levels, remaining stable in the following levels. This multiway approach can find specific patterns for gender differences that could be useful for promoting well-aimed actions to reduce the gap.

## 2. Gender gaps in mathematics

The issue of gender gap is a fact, which has considerable impact on the having and providing equal opportunities in terms of personal and professional life to females and males. The education sector has for some time now sought to address the issue of gender in the enrolment in scientific degree courses and the related impact on the job market.

The existing literature on gender gap, and especially on gender disparities in mathematics, is extensive and focuses particularly on examining differences in mean performance. This issue has been largely investigated in mathematics education from both qualitative and quantitative perspectives, by involving small studies or large-scale assessments (Cascella et al., 2020).

Several recent studies have investigated gender-related difference in mathematics more in detail. An understanding of whether or not gender-related differences in mathematical skills are present at different levels of the performance distribution has a pivotal role for defining educational, social and working policies. In fact, recognizing marked gender differences among the more successful students is important, as it would help in better understanding why girls, even though they are clever, decide not to pursue the study of STEM subjects. On the other hand, if large differences in the lower performance tails are identified, this would facilitate the implementation of learning activities aimed at tackling any shortcoming in the education of girls from the outset of their school life. The study of gender gap across performance distributions has roots in the works of Cleary (1992), Feingold (1992), Hedges and Friedman (1993), Halpern (1986) and Willingham and Cole (1997). In these works, the gender differences in performance were examined at different percentile points, taking into account the variability of the scores and examining the relationship of gender gaps to other factors affecting the gap.

It is beyond the scope of this paper to present a complete review on gender differences in mathematics. As in our analysis we focus on data coming from national standardised large-scale assessment, in the following we describe the main results in international and national literature dealing with large-scale assessments.

In this context, a common way to investigate gender-related differences in mathematics is to compute the average difference in performance, estimated through either the total test score or a model-based score. Examples of such models are the mixed coefficients multinomial logit model

(Adams et al., 1997), used in the OECD PISA survey, and the Rasch model (Rasch, 1960), used in the INVALSI assessments. A statistical test is usually conducted to understand whether the gap between females and males is significant.

The last OECD PISA survey conducted on 15-year-old students in 2018 showed some interesting results. In fact, despite boys have on average an advantage of five score points with respect to girls across OECD countries, the gap is not similar among all the 79 participating countries and economies (OECD, 2019b). In particular, boys significantly outperform girls in 32 countries/economies, while the opposite situation is observed in 14 countries/economies (OECD, 2019b). Unfortunately, Italy is one of the countries with the largest, keeping constant over time, mathematics gap in favour of boys (16 score points), and this difference is evident especially among the medium and highest performance levels (INVALSI, 2019b).

At national level, INVALSI data (INVALSI, 2017; 2018) confirm the international findings as boys significantly outperform girls in mathematics in all school grades in Italy. Moreover, the gap is evident already at primary school and it becomes rather large at grade 10 (about 10 score points in 2017). Also, mathematics gap is larger for top performers, e.g., students with the highest abilities, and differences by school type in higher secondary education are observed (INVALSI, 2017; 2018; 2019a).

The main results from the literature involving Italian data are summarised in the following. Ajello et al. (2018) focused on OECD PISA data showing that girls outperform boys in mathematics only when items have a high reading demand. De Simone (2013) analysed Italian TIMSS data by employing a pseudo-panel approach finding out that gender gap observed at grade 8 could be due to primary education (grade 4). Contini et al. (2017) analysed INVALSI data and show that, for students from grade 2 to 10, girls systematically underperform boys, even after controlling for individual and family background variables. Costanzo & Desimoni (2017)

evaluated the inequalities produced by gender and immigrant status in the educational outcomes along the performance distribution by using INVALSI primary school data. Matteucci & Mignani (2011) showed a significant gender gap at grade 8, but no inequalities by item type or domain. Furthermore, several authors focused on item-level analysis. Cascella (2015) and Cascella et al. (2020) investigated gender differences in specific items by using differential item functioning on Italian standardised data from INVALSI. A further study of mathematics gender gap of Italian grade 10 students is due to Ferretti & Giberti (2020), who focused on specific item content.

Despite international and national surveys highlight the importance of investigating gender differences in mathematics not only on average, but also by performance levels and on the entire ability scale distribution, the existing literature does not address these aspects in any great detail. For this reason, we believe that our proposal will be able to provide new insights in this research field.

## 3. Data and Measurement

### 3.1 Italian education system

The Italian education system is structured into three cycles. The first cycle includes primary and lower secondary education, and lasts 8 years. Primary school starts at 6 years of age and lasts 5 years (grades 1-5). Students with special needs are integrated into the mainstream education, and specialist support is provided. Lower secondary school starts at 11 years of age and lasts 3 years (grades 6-8).

The second cycle of education includes upper secondary school or regional vocational training system, and it is addressed to students starting from the age of 14. Upper secondary school lasts 5 years (grades 9-13) and the following pathways are possible: “licei”, that are more academic schools and, usually, prepare students to the university studies; technical schools, that prepare

students to work in agriculture, industry, commerce, administration and marketing; vocational schools, that offer vocational training for various jobs. The regional vocational training system offers 3 or 4-year courses organised by training agencies or upper secondary schools. In the past, at age 14, compulsory education was considered complete. Currently, compulsory education lasts until age 16.

The third education cycle includes post-secondary and higher education, i.e., university education, high level arts, music and dance education institutes, and specific higher education. Most university studies are organized into a bachelor degree (3 years) and a master degree (2 years) as a consequence of the Bologna Process which required a separation between first- and second-levels degrees.

In this paper, we consider data coming from the national large-scale tests administered by INVALSI during the first and the second cycle of education.

### 3.2 INVALSI design

Since the academic year 2008/2009, INVALSI has annually administered assessment tests to be taken by students at the end of the second and fifth years of primary school (grades 2 and 5), the third year of lower secondary school (grade 8) and the second year of upper secondary school (grade 10). Until the academic year 2012/2013, tests for grade 6 were administered. Moreover, from the academic year 2018/2019, tests have been administered also to 13 grade students at the end of upper secondary school.

The assessment entails answering a test about the Italian language (reading comprehension and grammar) and a test about mathematics, which must be completed by all enrolled students. The tests are administered to the whole student population, around 500,000 students for each grade. INVALSI also builds a random sample of around 25,000 units for each grade.

The sampling procedure is a two-stage with geographic stratification at the first stage. The units of the first stage are the schools and the units of the second stage are the classes. At the first stage, the stratification is based on the regional area. For a specific stratum, the units are selected with no re-entry and the inclusion probability of a school is proportional to the number of students. At the second stage, for each sampled school, a sample of two classes is randomly selected. Each pupil and each class of the stratum have the same probability of being selected in the sample, regardless of the size of the school selected in the first stage. For further details see Falorsi et al. (2019).

The Institute processes the data and produces a report, the relevant results of which are then sent to each school. For research purposes, it also provides the micro data (in anonymous form) related to each student, as well as certain information on the performance and socio-economic characteristics of the student and his or her family.

### 3.3 Participants

In this paper, the analysis is based on data obtained from the INVALSI mathematics tests administered in the 2016-2017 academic year to students in grades 5, 8 and 10 (INVALSI, 2017). These tests have been administered in May 2017, about one month before the summer holidays.

Since the academic year 2017/2018, INVALSI has administered computer based tests for grades 8 and 10 and also for grade 13 starting from the next year. In our analyses, we have considered the previous academic year 2016/2017 so that the test administration conditions are the same for all grades and the results allow for a better comparison.

In this paper, we used a simple random sub-sample of 5,000 pupils from the national INVALSI sample to allow the implementation of the methodological tools that require a great deal of computational efforts. Checks were carried out ensure the sample was representative in terms of the gender structure.

### 3.4 The mathematics tests

The mathematics tests are provided in paper and pencil form and, consistently with international research, contain items in 4 different content domains: Numbers, Space and Figures, Data and Predictions, and Relations and Functions.

The types of items include multiple choice, open but with only one possible answer, or open with or without the requirement to justify the answer given. The answers are corrected by experts and the results codified in binary 0-1 (incorrect or correct). Table 1 shows the structure of the test questionnaire for each grade.

| | Grade 5 (75 min.) | Grade 8 (75 min.) | Grade 10 (90 min.) |
|---|---|---|---|
| Content category | Number of items | Number of items | Number of items |
| Numbers | 12 (26%) | 10 (20%) | 18 (34%) |
| Space and Figures | 10 (22%) | 13 (26%) | 9 (17%) |
| Data and Predictions | 14 (30%) | 12 (24%) | 11 (21%) |
| Relations and Functions | 10 (22%) | 15 (30%) | 15 (28%) |
| Item type | | | |
| True/false | 9 (20%) | 14 (28%) | 18 (34%) |
| Multiple Choice | 13 (28%) | 14 (28%) | 14 (26%) |
| Open answer (unique answer) | 22 (48%) | 19 (38%) | 19 (36%) |
| Open long answer (argumented or not) | 2 (4%) | 3 (6%) | 2 (4%) |
| **Total** | **46** | **50** | **53** |

**TABLE 1**. *Description of the tests by content domain, item type and test time for the three grades (column percentages for item domain and type in brackets).*

The tests show an increase in the test length as the grades increase. With regard to the 4 different contents, the item distribution is varying depending on the level at which, according to the guidelines given by the Ministry of Education, these areas are to be studied.

### 3.5 The INVALSI measurement

As outcome variable, INVALSI calculates a raw score and, using the Rasch model, an ability score on a continuous scale centred at zero and variance one, namely the Rasch score.

The Rasch measurement model, developed by George Rasch (Rasch, 1960), has been widely used to measure ability using the answers of a test. It is a popular approach for modelling the probabilistic relationship between responses to test items and individual abilities. If the item response is binary (correct/wrong), the probability of giving a correct answer to a generic item *j*, is expressed by a logistic model, as follows:

$$P_j(\theta) = \frac{\exp(\theta - b_j)}{1 + \exp(\theta - b_j)}, \quad (1)$$

where $\theta$ is the ability and $b_j$ is the difficulty of item *j*. The difficulty parameter is equivalent to the point on the ability scale where the probability of answering item *j* correctly is equal to 0.5. The higher the difficulty parameter, the higher the ability level required to answer the item correctly.

The characteristics of the Rasch model measurement are: 1) it transforms the students' probability of solving items to log odds and logits as the units in the Rasch model; 2) items and students are placed on the same interval scale, allowing for the prediction of the participants' responses to a particular item; 3) it provides information to examine whether each item fits within

the underlying trait (ability). The 'ability' variable is defined in the (-∞, +∞) range and follows a standard normal distribution at the population level.

## 4. Methods

The first step of the data analysis entailed a descriptive analysis of the percentage of correct answers distinguished between boys and girls. The performance averages of boys and girls were compared based on the estimated Rasch score.

The data were then compared along the entire performance distribution using linear quantile regression model. In this model, the Rasch score is the dependent variable while the gender is the independent variable. By using a quantile approach, it is possible to assess if the gender effect is different according to the student's level of ability.

In the last step, a classification procedure is applied to determine groups with similar performance levels, in order to identify lower and higher performance characteristics in relation to gender. Latent class analysis with gender as explanatory variable was used.

### 4.1 Quantile regression

Quantile regression, as introduced by Koenker and Bassett (1978), may be considered as an extension of classical least squares estimation of conditional mean models, to the estimation of a set of conditional quantile functions.

Quantile regression is used to deepen the study of the linear relation between a dependent variable $y$ and a set of covariates ($x$) among the quantiles ($q$) of the dependent variable (see, e.g., Davino et al., 2013; Koenker, 2010). It is a regression technique, which allows to focus on the effects that the explanatory variables have on the entire conditional distribution of the dependent variable, namely it takes into account that this effect can be different for students with different levels of ability.

Classical linear regression techniques summarize the average relationship between a set of explanatory variables and the outcome variable based on the conditional mean function E($y|x$). This provides only a partial view of the relationship, as we might be interested in describing the relationship at different points in the conditional distribution of $y$. Quantile regression provides that capability.

The quantile regression model is described by the following equation:

$$y_i = \boldsymbol{x'}_i \boldsymbol{\beta}_q + \varepsilon_i, \qquad (2)$$

where $\boldsymbol{\beta}_q$ is the vector of regression coefficient associated with the $q$-th quantile, with $i$=1,..,$n$ subjects. A quantile regression parameter estimates the change in a specified quantile $q$ of the dependent variable produced by one unit change in the independent variable.

To evaluate the impact of the independent variable along the entire distribution of the dependent variable, the comparison among the estimated regression coefficients could be useful. A hypothesis test can be implemented to verify at which quantile the difference is significant.

**4.2 Latent class analysis**

Latent class analysis (LCA) is used for clustering purposes, i.e., it allows for the allocation of individuals into homogeneous groups based on the item responses. As a result, in addition to the classification and the heuristic interpretation of the same, it is possible to estimate the probability of belonging to a particular latent class (group).

With regard to certain socio-behavioural phenomena, the existence of subgroups of individuals with common traits can be hypothesised, although not observed directly. In the case at hand, the existence of subgroups of students who respond in a similar way to the items can be assumed.

The idea of measuring latent attitudinal variables based on dichotomous items was initially proposed by Lazarfeld & Henry (1968) and was subsequently formalised by Goodman (1974).

In latent class analysis, each class is characterised by a set of conditional probabilities, i.e., the probabilities that the observed variables assume a given value (e.g., 0-1) within a certain class. The statistical units are assigned to classes based on their class-membership probability, thus creating groups that are mutually exclusive.

With this method, therefore, it is possible to subdivide a heterogeneous population into subgroups that are internally homogeneous but heterogeneously distinct from one another, without making any restrictive assumptions about the data, which are thus less subject to potential bias due to any incongruity with the initial hypotheses. Furthermore, the addition of covariates to the analysis allows for an effective description of the groups.

Traditional latent class analysis aims to identify the lowest number of C latent classes, in order to explain the observed associations among the manifest variables (the observed item responses) based on the data. With this method, it is possible to estimate the prior probabilities of latent class membership, the probabilities of a given item response, conditional to the class membership (conditional probabilities) and the probabilities of belonging to a certain latent class, given the item responses (posterior probabilities), calculated after the estimation of the model parameters.

An important extension of the traditional LCA is the latent class regression model (Dayton & Macready, 1988; Hagenaars & McCutcheon, 2002) where it is possible to include covariates to predict the latent class membership. It is supposed that an individual does not belong to a certain class solely based on the answers given, but also on personal characteristics. A generalised multinomial logit link function is used to specify the effects of the covariates on the prior probabilities, as follows:

$$\ln\left(\frac{p_{2i}}{p_{1i}}\right) = \boldsymbol{X}_i\boldsymbol{\beta}_2$$
$$\ln\left(\frac{p_{3i}}{p_{1i}}\right) = \boldsymbol{X}_i\boldsymbol{\beta}_3$$
$$\ldots \qquad (3)$$
$$\ln\left(\frac{p_{Ci}}{p_{1i}}\right) = \boldsymbol{X}_i\boldsymbol{\beta}_C,$$

where $i$=1,…,$n$ subjects, $p_c$, with $c$=1,..,$C$ classes, is the prior probability of belonging to class $c$, $\mathbf{X}_i$ is the vector of covariates for the individual $i$ and $\boldsymbol{\beta}_c$ is the vector of regression coefficients corresponding to the latent class $C$. By including a single covariate, an intercept term and a regression coefficient are estimated for each class. Class 1 is usually taken as the "reference class" (see Linzer & Lewis, 2011).

By including the gender covariate in a latent class model, it is possible to model the membership of the high and lower classes, with reference to the average class, and to evaluate the impact of gender on the membership of a certain class. The aim of this analysis is to find an alternative method of verifying whether gender has an effect and at which performance levels.

## 5. Results

In this section, we present the main results of the descriptive analyses, the quantile regression model and the latent class regression model by school grades also including a description of groups by content domains. The analyses were performed using the R software (R Core Team, 2019), specifically, the quantreg package for quantile regression (Koenker, 2019) and the poLCA package for the latent class regression analysis (Linzer & Lewis, 2011).

It is important to specify that the tests were subject to an analysis to detect the presence of any differential item function (DIF), with regard to gender in order to ensure accurate comparison. For each questionnaire, the estimated DIF was less than 7% and, therefore, it was deemed that the questionnaire was more than satisfactory in order to proceed.

### 5.1 Grade 5 results

Grade 5 represents the last year of primary school and it is an important stage at which an initial assessment of the learning achieved by the students should be performed.

**Comparison of overall performance between boys and girls**. An initial straightforward analysis was performed comparing performance in each subject area with regard to gender.

| | Total | Male | Female | |
|---|---|---|---|---|
| Total | 53.96 | 55.72 | 52.10 | *** |
| | (20.31) | (20.23) | (20.24) | |
| Numbers | 54.74 | 57.43 | 51.98 | *** |
| | (25.50) | (25.30) | (25.42) | |
| Data and Predictions | 62.44 | 63.48 | 61.38 | ** |
| | (29.02) | (23.90) | (24.12) | |
| Space and Figures | 53.24 | 53.75 | 52.72 | n.s. |
| | (21.08) | (20.89) | (21.27) | |
| Relations and Functions | 45.37 | 48.03 | 42.64 | *** |
| | (25.81) | (25.83) | (25.50) | |

**TABLE 2**. *Percentage of correct answers by gender and content domain, grade 5 (standard deviations are in brackets, *p-value <0.05, **p value <0.01, ***p-value <0.001, n.s. not-significant).*

The data in Table 2 show the percentages of correct answers and the level of significance of the difference between boys and girls (M-F). The overall performance level is good with the 53.96%

of correct answers and results above 52% for both boys and girls. The content domain Relations and Functions looks like the most difficult, while Data and Predictions is the easiest. The percentages for boys are higher than those of girls, for all the content domains. However, the answer percentages vary according to variations in the content in the same way for both boys and girls. With regard to the gap between boys and girls, the Space and Figures area is the only one in which there is no significant difference in terms of performance. It is important to emphasise this result as it contradicts the findings in much of the literature, according to which boys usually outperform girls in areas related to space and geometry in general (Battista, 1990; Erdoğan et al., 2011; Nemeth, 2007). Different views have been proposed to explain boys and girls differences, related to both cognitive factors and environmental factors, such as activities in which boys and girls are engaged in their daily life (see, e.g., Arnup et al., 2013). There is a strong possibility that this is also due to how the questions are formulated, but it nonetheless merits further investigation. In general, the variability in the answers given for each area is high and relatively similar for both boys and girls. This result confirms the performance heterogeneity and the need, therefore, for a more detailed analysis conducted using quantile regression and latent class models.

Subsequently, the different performance levels were verified at a general level for boys and girls, with the score obtained through the Rasch model considered as the response variable. A test on the mean was therefore performed on this variable. The mean value for boys is higher and is equal to 0.068 (s.d. 1.098), while the value for girls is -0.119 (s.d. 1.090). The difference is significant at the 0.001 level.

**Quantile regression model.** The analysis then proceeded with an examination of the differences along the entire performance scale. The following graph shows the percentile distribution of ability for boys and girls.

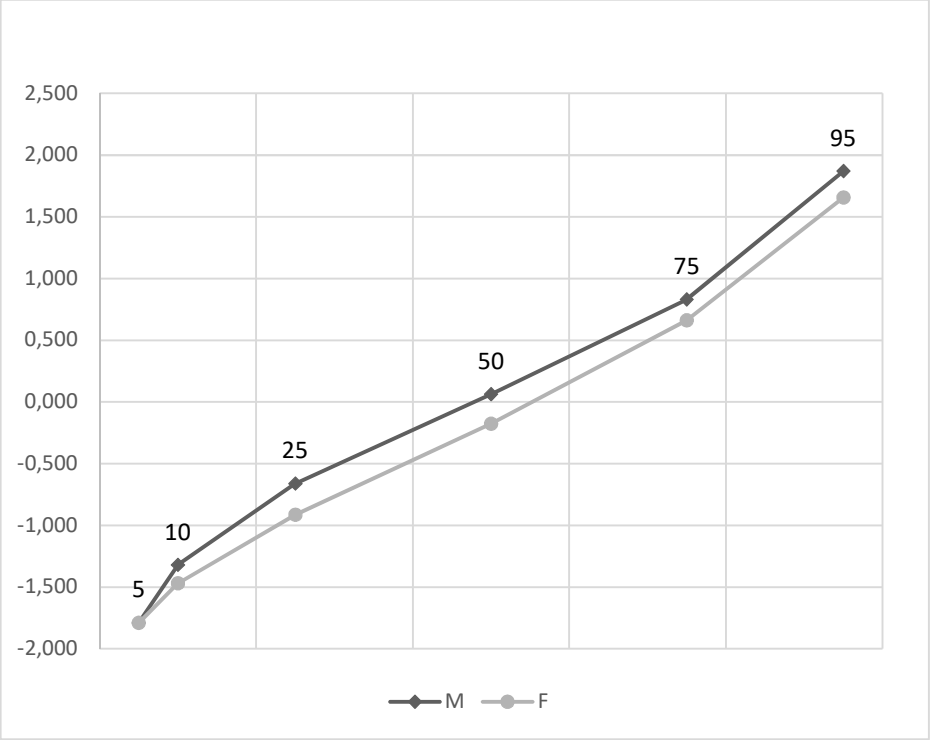


**FIGURE 1**. *Percentile distribution of performance based on Rasch model by gender, grade 5.*

The graph shows that boys (M) outperform than girls (F) along the entire scale, with the exception of the lower percentiles. In order to study the behaviour along the entire distribution scale, a quantile linear regression model approach was used.

| Quantile | Intercept (*p*-value) | Coefficient (*p*-value) | *p*-value of the difference between two consecutive coefficients |
|---|---|---|---|
| 0.05 | -1.791 *** | 0.000 *n.s.* | --- |
| | (0.000) | (1.000) | |
| 0.10 | -1.321 *** | -0.140 ** | 0.000 |
| | (0.000) | (0.002) | |
| 0.25 | -0.661 *** | -0.252 *** | 0.020 |
| | (0.000) | (0.000) | |
| 0.50 | 0.062 *** | -0.240 *** | 0.775 |
| | (0.005) | (0.000) | |
| 0.75 | 0.831 *** | -0.252 *** | 0.750 |
| | (0.000) | (0.000) | |
| 0.95 | 1.870 *** | -0.215 *** | 0.484 |
| | (0.000) | (0.000) | |

**TABLE 3**. *Estimated model parameters for six quantiles, with Male as reference category, grade 5 (*p-value <0.05, **p value <0.01, ***p-value <0.001, n.s. not-significant*).

Table 3 shows the regression coefficients of the models for the six values of percentiles, where, for the gender covariate, the reference is the male category. The results highlight that the lowest performance levels (5$^{th}$ percentile) are similar for boys and girls. However, as of the tenth percentile, the differences are significant and, since the value is negative, the fact of being female rather than male determines on average a reduction in the performance level (Rasch score) equal to the value indicated by the regression coefficient. The most notable impact is seen at the 25$^{th}$ and

the 75th percentiles. In the last column, the *p*-values for comparing consecutive regression coefficients are included. In particular, a significance test is conducted to compare the difference between each coefficient and the previous one, so that five *p*-values are reported. Considering this comparison, the test results confirm that, starting from the 25th percentile, the impact of gender does not significantly change over percentiles.

**Latent class analysis**. A latent class analysis was then performed with the addition of the gender covariate. The number of classes was chosen based on a comparison among three different models with two, three and four classes. Table 4 shows the values of the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), classical indexes to be used for choosing the best-fitting model among several competing nested models.

| Number of latent classes | AIC | BIC |
|---|---|---|
| Two | 215843.1 | 216379.7 |
| Three | 212123.3 | 212933.7 |
| Four | 211264.6 | 212337.9 |

**TABLE 4.** *AIC and BIC by number of latent classes, grade 5.*

As can be noticed from Table 4, both AIC and BIC decrease as the number of classes increases and the decrease is more evident moving from two to three groups. Despite the lowest values for AIC and BIC are associated to the four-class solution, the model with three classes shows more interpretable results compared to the model with four classes, on the basis of the conditional probabilities of a correct answer for each item within each class. In fact, unlike the model with four classes, in the three-class model all the conditional item probabilities show similar behaviour inside

the same group, allowing an easy interpretation of performance in mathematics. For this reason, we chose the model with three latent classes, which seems to be the adequate solution to identify reasonable groups of students with similar level of performance.

The three classes can be interpreted as follows: "Low" (lowest performance), characterised by the lowest probability of a correct answer to all the items compared with the other two classes; "Medium" (medium performance), with higher probabilities than the previous class, but lower than that of the last class; and "High" (highest performance), with the highest conditional probabilities out of all three classes. With this model, therefore, it was possible to estimate the probability of membership of each class (prior probabilities), also by gender. These probabilities are shown in Table 5.

| | Low | Medium | High |
|---|---|---|---|
| Male | 0.27 | 0.49 | 0.24 |
| Female | 0.30 | 0.54 | 0.16 |
| *p*-value | 0.034* | 0.003** | 0.000*** |

**TABLE 5**. *Probability of class membership by class and gender, grade 5 (*p-value <0.05, **p value <0.01, ***p-value <0.001, n.s. not-significant*).

For both genders, the Medium class has the highest prior probability while the High class has the lowest. The greatest difference is once again found in relation to the highest performance level: the prior probability of belonging to the High class is 0.24 for males and only 0.16 for females. The last result obtained from the analysis is a further confirmation of the comparatively lower performance of girls. We applied a significance test to assess if the probability of belonging to a

specific class is significantly different for males and females. The results are significant for all the three classes and especially for the High class ($p$-value <0.001).

Table 6 shows the estimated parameters of model (3) applied to the three classes with the gender covariate. The Medium class was taken as the reference category. Therefore, we have two different logistic models: model 1, where the High class is compared with the reference, and model 2, where the Low class is compared with the reference. The gender covariate is introduced with "male" as reference category, i.e., the regression coefficient will evaluate the effect of being a female (X=1) rather than a male (X=0). For making the interpretation of the results simpler, Table 6 reports only the odds for males, the odds for females and the odds ratio.

| | High vs. Medium (model 1) | Low vs. Medium (model 2) |
|---|---|---|
| Odds Male | 0.78 | 0.54 |
| Odds Female | 0.49 | 0.55 |
| Odds ratio Female/Male | 0.63*** | 1.01 (n.s) |

**TABLE 6**. *Estimated parameters in the two logistic regression models obtained by the latent class analysis, grade 5 (*p-value <0.05, **p value <0.01, ***p-value <0.001, n.s. not-significant*).

The values in Table 6 can be interpreted as follows. The Odds Male is the estimated odds (probability of being in a certain class divided by the probability of being in the reference class) for males. The Odds Female has an analogous interpretation, but for females. Then, the odds ratio (OR) is calculated by dividing the two odds. When the OR is equal to one, no gender effect is observed on the probability of being in a certain latent class, in comparison to the reference class.

In model 1, the odds of passing from the Medium to the High class is equal to 0.78 for boys and 0.49 for girls, denoting a gender gap in favour of boys. In fact, the OR is equal to 0.63, significantly different from 1, meaning that the transition from the Medium to the High latent class is more likely for males than females. In model 2, the estimated odds of passing from the Medium to the Low class is equal to 0.54 for males and 0.55 for females. In this case, the OR is not significantly different from 1, meaning that there is no gender effect in the probability of being in the Low class in comparison to the Medium class. A gender-related difference is once again confirmed at the higher performance levels.

Based on this latent class analysis, we were thus able to estimate the posterior probability of the students' membership of the three classes. The classification obtained served as a basis for a more in-depth analysis of the groups with regard to performance in the content domains. Table 7 shows the percentage of correct answers given by boys and girls for the different contents in relation to the three classes.

Overall, the differences between boys and girls in each group are not significant as to indicate similar behaviours. An examination of the different contents reveals some differences in favour of the girls. In general, boys and girls in all groups find Relations and Functions the hardest subject area, with the exception of the higher-performing students, where, for boys, the hardest topic is Space and Figures, a content which is significantly more favourable for girls. In the Medium group, the differences are significant for all domains, except for Relations and Functions. It should be noted that in both Data and Predictions and Space and Figures subject contents, girls achieve higher results than boys, while boys perform higher on Numbers items. Within the High class, girls outperform boys in Geometry, a content within the Space and Figures domain, in which boys typically outperform girls.

| | Low | | | Medium | | | High | | |
|---|---|---|---|---|---|---|---|---|---|
| | M | F | Diff | M | F | Diff | M | F | Diff |
| Total | 28.35 | 28.59 | n.s. | 53.66 | 53.49 | n.s. | 77.65 | 77.86 | n.s. |
| Numbers | 25.80 | 25.92 | n.s. | 55.93 | 53.05 | *** | 81.69 | 81.05 | n.s. |
| Space and Figures | 33.83 | 35.72 | n.s. | 64.21 | 66.28 | ** | 83.57 | 83.63 | n.s. |
| Data and Predictions | 35.77 | 35.83 | n.s. | 50.81 | 52.43 | * | 70.08 | 73.17 | *** |
| Relations and Functions | 18.75 | 17.59 | n.s. | 43.41 | 42.11 | n.s. | 74.48 | 73.12 | n.s. |

**TABLE 7**. *Percentages of correct answer by content domain, gender and latent class, and test for the mean difference, grade 5 (*p-value <0.05, **p value <0.01, ***p-value <0.001, n.s. not-significant).*

Considering all the results for grade 5, the gap is already established and it is significant from low-medium performance levels and it remains significant with a stable impact for the following ability levels. The LCA groups confirm the main findings.

### 5.2 Grade 8 results

Grade 8 is the last year of lower secondary school, a key point in time for determining a student's subsequent course of study. Schools provide students with career guidance in order to help them decide which course of study to pursue. The INVALSI test therefore offers further food for thought as each student receives an individual score, whereas at the preceding and subsequent levels, the results are provided in aggregate form at the class level.

**Comparison of overall performance between boys and girls**. Table 8 shows the average percentage of correct answers to the items classified by content and distinguished by gender.

| | Total | Male | Female | |
|---|---|---|---|---|
| Total | 53.73 | 55.08 | 52.41 | *** |
| | (20.16) | (20.12) | (19.72) | |
| Numbers | 46.92 | 48.50 | 45.29 | *** |
| | (23.73) | (23.61) | (23.74) | |
| Data and Predictions | 67.45 | 68.89 | 65.96 | *** |
| | (22,25) | (22.14) | (22.26) | |
| Space and Figures | 50.32 | 50.47 | 50.18 | n.s. |
| | (22.19) | (22.60) | (21.77) | |
| Relations and Functions | 51.79 | 53.69 | 49.83 | **** |
| | (25.78) | (25.62) | (19.72) | |

**TABLE 8.** *Percentage of correct answers by gender and content domain, grade 8 (standard deviations are in brackets, *p-value <0.05, **p value <0.01, ***p-value <0.001, n.s. not-significant).*

Overall performance levels are higher than 50% of correct responses. The percentages for boys are higher than those of girls, for all content domains. However, the answer percentages vary according to variations in the contents in the same way for both boys and girls. Both genders find Data and Predictions to be the easiest subject area while Numbers is the hardest, the latter result being in contrast to the findings related to grade 5. However, it is important to underline that the comparison among grades is not properly correct because the student cohorts are different. Anyway, the comments are a way to suggest possible insights to be investigated if longitudinal data will be available in the future. It should be pointed out that, in the Relations and Functions domain, the variability in the performance of boys is particularly high, while in that of girls is the

lowest. The only area in which there are no significant differences is Space and Figures, which is confirmed therefore as the most balanced area.

Looking at the Rasch score, the mean value for boys is 0.068 (s.d. 1.15) compared with -0.082 (s.d. 1.13) for girls, a significant difference at the level of 0.001.

**Quantile regression model**. Figure 2 shows the percentile distribution by gender.

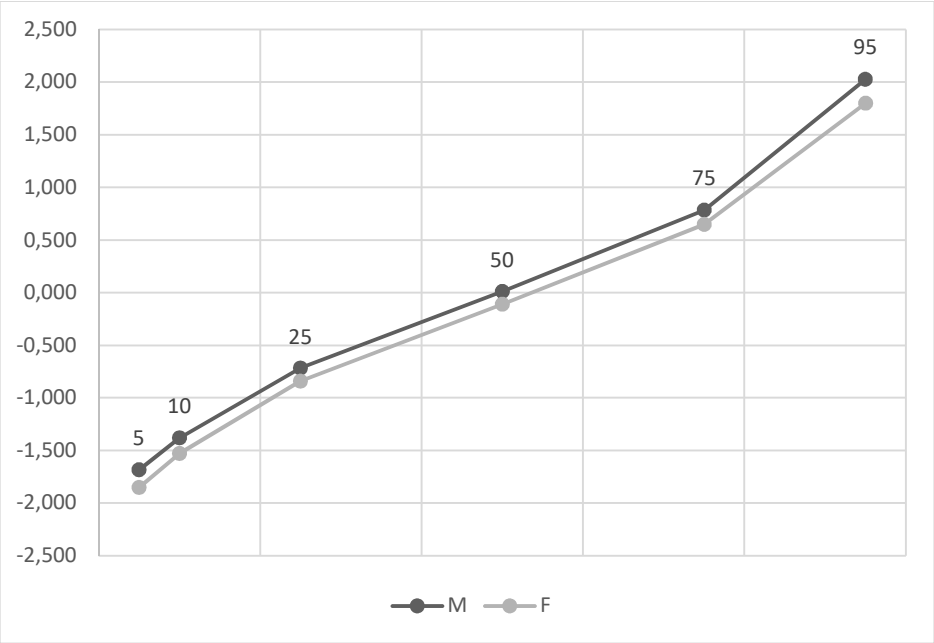


**FIGURE 2.** *Percentile distribution of performance based on Rasch model by gender, grade 8.*

The performance levels of boys are consistently higher at all ability levels, including the lowest.

The results in Table 9 shows that the differences are significant along the entire distribution curve, starting from the 5$^{th}$ percentile, with a negative value. The differences among consecutive coefficients are all not significant, meaning that the gap is stable over percentiles, i.e., the impact of gender is similar along the entire performance distribution.

| Quantile | Intercept (*p*-value) | Coefficient (*p*-value) | *p*-value of the difference between two consecutive coefficients |
|---|---|---|---|
| 0.05 | -1.685 *** | -0.168 * | --- |
| | (0.000) | (0.012) | |
| 0.10 | -1.380 *** | -0.148 ** | 0.733 |
| | (0.000) | (0.002) | |
| 0.25 | -0.661 *** | -0.252 *** | 0.577 |
| | (0.000) | (0.000) | |
| 0.50 | 0.062 | -0.240 *** | 0.902 |
| | (0.004) | (0.000) | |
| 0.75 | 0.831*** | -0.252 *** | 0.628 |
| | (0.000) | (0.000) | |
| 0.95 | 2.026 *** | -0.228 *** | 0.122 |
| | (0.000) | (0.000) | |

**TABLE 9**. *Estimated model parameters for six quantiles, with Male as reference category, grade 8 (*p-value <0.05, **p value <0.01, ***p-value <0.001, n.s. not-significant*).

**Latent class analysis**. With the same approach adopted for grade 5, three classes were chosen: Low, Medium and High.

| | Low | Medium | High |
|---|---|---|---|
| Male | 0.33 | 0.45 | 0.22 |
| Female | 0.39 | 0.43 | 0.17 |

| | | | |
|---|---|---|---|
| *p*-value | 0.000*** | 0.077 | 0.000*** |

**TABLE 10.** *Probability of class membership by class and gender, grade 8 (*p-value <0.05, **p value <0.01, ***p-value <0.001, n.s. not-significant*).

As reported in Table 10, for both genders, membership of the Medium class is most probable, while membership of the High class is the least probable. Again, the prior probability of belonging to the High class is 0.22 for males and only 0.17 for females. We can observe that the difference for the Medium class is the only not significant.

| | High vs. Medium (model 1) | Low vs. Medium (model 2) |
|---|---|---|
| Odds Male | 0.61 | 0.61 |
| Odds Female | 0.50 | 0.74 |
| Odds ratio Female/Male | 0.82*** | 1.21*** |

**TABLE 11**. *Estimated parameters in the two logistic regression models obtained by the latent class analysis, grade 8 (*p-value <0.05, **p value <0.01, ***p-value <0.001, n.s. not-significant*).

Looking at Table 11, in model 1, the odds for boys is equal to 0.61, meaning that being in the Medium class is more likely than being in the High class. Analogously, the estimated odds for girls is 0.5. The OR is significant and equal to 0.82: for females, the probability of being in the High

class in comparison to the Medium class is 0.8 times the same rate for males. Again, girls are less favoured than boys in the transition from the Medium to the High class. In model 2, the odds for boys is equal to 0.61 while for girls it is 0.74. Even if the probability of belonging to the Low class is lower than the one of belonging to the Medium class, both for boys and girls, this rate is higher for girls. Here, there is a significant effect of gender on the probability of moving from the Medium to the Low class. In particular, the OR is equal to 1.21, higher than 1, meaning that being a female instead of a male increases the probability of being in the Low class compared to the Medium one. At this education stage, therefore, the disadvantage of girls in relation to the highest performance, which is more evident than at grade 5, is confirmed.

Table 12 shows the percentage of correct answers given by boys and girls belonging to the three latent classes with respect to the different item domains.

| | Low | | | Medium | | | High | | |
|---|---|---|---|---|---|---|---|---|---|
| | M | F | Diff | M | F | Diff | M | F | Diff |
| Total | 29.38 | 29.46 | n.s. | 52.82 | 53.21 | n.s. | 73.77 | 72.62 | n.s. |
| Numbers | 23.85 | 23.00 | n.s | 47.81 | 47.96 | n.s. | 74.52 | 73.74 | n.s. |
| Space and Figures | 45.09 | 44.42 | n.s. | 71.04 | 71.41 | n.s. | 89.32 | 87.61 | n.s. |
| Data and Predictions | 28.36 | 30.97 | *** | 49.19 | 51.55 | *** | 74.90 | 76.59 | * |
| Relations and Functions | 24.48 | 23.51 | n.s. | 54.44 | 53.69 | n.s. | 81.90 | 81.97 | n.s. |

**TABLE 12**. *Percentages of correct answer by content domain, gender and latent class and test for the mean difference, grade 8 (*p-value <0.05, **p value <0.01, ***p-value <0.001, n.s. not-significant).*

At this grade, there are generally no differences between boys and girls within each group and both boys and girls in all groups now find Numbers to be the hardest subject area. Breaking down the performance results, the finding is confirmed but there are no significant differences except with regard to Space and Figures, where girls outperform boys in all groups.

To sum up, at grade 8, the gender gap is significant already for very low ability levels, showing successively a stable impact from the medium levels. These results are confirmed by the classification made by using LCA.

### 4.3 Grade 10 results

Grade 10 represent the last year of compulsory schooling, where the mathematics contents is the same for all types of schools.

**Comparison of overall performance between boys and girls**. Table 13 shows the average percentage of correct answers to the items classified by subject area and distinguished by gender.

| | Total | Male | Female | |
|---|---|---|---|---|
| Total | 47.29 | 48.50 | 45.95 | *** |
| | (21.98) | (22.61) | (21.18) | |
| Numbers | 49.18 | 50.37 | 47.86 | *** |
| | (23.20) | (23.81) | (22.43) | |
| Data and Predictions | 53.75 | 55.54 | 51.76 | *** |
| | (24.37) | (24.76) | (23.77) | |
| Space and Figures | 40.313 | 41.28 | 39.45 | * |

| | | | | |
|---|---|---|---|---|
| | (26.21) | (26,50) | (25.85) | |
| Relations and Functions | 45.48 | 46.60 | 44.24 | ** |
| | (28.05) | (28.93) | (26.98) | |

**TABLE 13**. *Percentage of correct answers by gender and content domain, grade 10 (standard deviations are in brackets, *p-value <0.05, **p value <0.01, ***p-value <0.001, n.s. not-significant).*

Overall, we observe a slightly decreasing performance with most percentages below 50% for both genders and, in any case, lower than that of previous education stages. There is, however, a higher degree of variability with differences in performance starting to become more evident. The percentages for boys are higher overall in all the contents. Both genders find Data and Predictions to be the easiest subject area while, in contrast to the findings related to grade 8, Space and Figures is the hardest.

Looking at the Rasch score, the mean value for boys is 0.11 (s.d. 1.21) compared with -0.04 (s.d. 1.12) for girls, a difference that has a significance level of 0.001 also for this school level.

**Quantile regression**. Figure 3 shows the performance percentile distribution by gender.

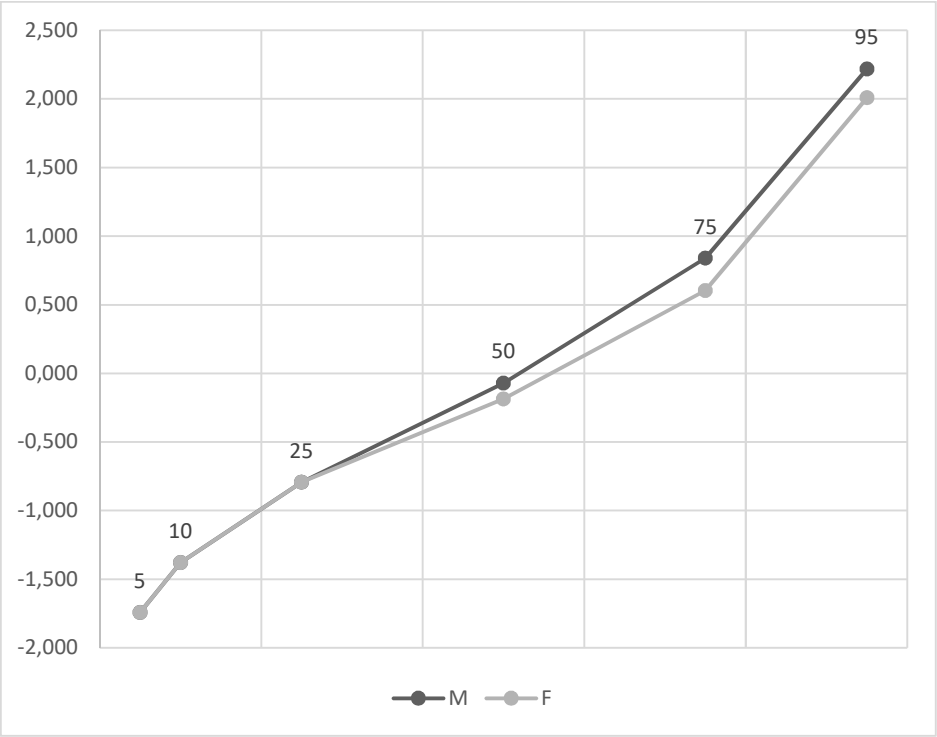


**FIGURE 3**. *Percentile distribution of performance based on Rasch model by gender, grade10.*

The graph shows that, at this stage, the differences between boys and girls only start to appear around the medium levels in favour of boys and that the values are subsequently maintained.

From the results of the quantile regression model in Table 14, it can be inferred that the differences start to become significant at the medium performance levels (50$^{th}$ percentile) and are particularly marked in the subsequent percentile (75$^{th}$). The difference however reduces at the higher-performing levels (coefficient equal to -0.208 at the 95$^{th}$ percentile). It thus appears that, among students with low ability, there is no significant difference between boys and girls. At the medium level, however, boys have higher performance than girls. The comparison of consecutive regression coefficients confirm that the impact of gender is not stable along the whole distribution, as the differences are significant from the 25$^{th}$ to the 50$^{th}$ percentiles and for all the following comparisons.

This finding differs from that concerning previous education level and highlights an important change at grade 10, due probably to the type of choices made with regard to future study.

| Quantile | Intercept (*p*-value) | Coefficient (*p*-value) | *p*-value of the difference between two consecutive coefficients |
|---|---|---|---|
| 0.05 | -1.742 *** | 0.000 | --- |
| | (0.000) | (1.000) | |
| 0.10 | -1.378 *** | -0.001 | 1.000 |
| | (0.000) | (0.980) | |
| 0.25 | -0.793 *** | -0.001 | 1.000 |
| | (0.000) | (0.970) | |
| 0.50 | -0.072 * | -0.114 ** | 0.002 |
| | (0.013) | (0.007) | |
| 0.75 | 0.960 *** | -0.356 *** | 0.000 |
| | (0.000) | (0.000) | |
| 0.95 | 2.216 *** | -0.208 ** | 0.060 |
| | (0.000) | (0.010) | |

**TABLE 14**. *Estimated model parameters for six quantiles, with Male as reference category, grade 10 (*p-value <0.05, **p value <0.01, ***p-value <0.001, n.s. not-significant*).

**Latent class analysis**. As in the analysis of previous grades, three classes are evaluated for this grade as well. Table 15 shows the probability of class membership also based on gender.

| | Low | Medium | High |
|---|---|---|---|
| Male | 0.34 | 0.44 | 0.21 |
| Female | 0.34 | 0.49 | 0.17 |
| *p*-value | 1 | 0.000*** | 0.001** |

**TABLE 15**. *Probability of class membership by class and gender, grade 10 (*p-value <0.05, **p value <0.01, ***p-value <0.001, n.s. not-significant*).

A breakdown of the results by gender shows that the probability of membership of the High class for girls is lower than for boys, albeit with a smaller difference between the two. It is interesting to note that the probability of membership of the Low class is practically the same for both genders, as confirmed by the not significant test.

Table 16 shows the parameters of the logistic regression models applied to the three classes with the gender covariate. In model 1, the odds of the probability of being in the High class rather in the Medium class is equal to 0.68 for boys and 0.49 for girls, meaning that the latter are not favoured in the transition from the Medium to the High class. In fact, the OR is equal to 0.71: for females, the probability of belonging to the High class relatively to the Medium class is only about 0.7 times the same rate for males. In model 2, the gender covariate has no significant effect in the odds of belonging to the Low class rather to the Medium one.

| | High vs. Medium (model 1) | Low vs. Medium (model 2) |
|---|---|---|
| Odds Male | 0.68 | 0.86 |
| Odds Female | 0.49 | 0.78 |
| Odds ratio Female/Male | 0.71*** | 0.90 (n.s.) |

**TABLE 16**. *Estimated parameters in the two logistic regression models obtained by the latent class analysis, grade 10 (*p-value <0.05, **p value <0.01, ***p-value <0.001, n.s. not-significant*).

| | Low | | | Medium | | | High | | |
|---|---|---|---|---|---|---|---|---|---|
| | M | F | Diff | M | F | Diff | M | F | Diff |
| Total | 24.51 | 24.00 | n.s. | 48.81 | 47.90 | * | 78.61 | 77.14 | ** |
| Numbers | 28.41 | 27.55 | n.s. | 50.25 | 49.87 | n.s. | 78.53 | 76.25 | ** |
| Space and Figures | 31.83 | 30.52 | n.s. | 59.31 | 56.02 | *** | 80.18 | 76.98 | *** |
| Data and Predictions | 18.55 | 17.02 | * | 39.37 | 40.56 | n.s. | 72.32 | 73.11 | n.s. |
| Relations and Functions | 18.60 | 19.92 | * | 46.13 | 44.84 | n.s. | 82.98 | 81.96 | n.s. |

**TABLE 17**. *Percentages of correct answer by content domain, gender and latent class and test for the mean difference, grade 10 (*p-value <0.05, **p value <0.01, ***p-value <0.001, n.s. not-significant).*

Table 17 shows the percentage of correct answers given by boys and girls in the different contents in relation to the three classes. At grade 10, things change slightly compared with previous school grades. Indeed, it is only in the Low class that the performance of boys and girls is much similar, whereas in the other two groups, there is a significant difference at the overall level. With regard to the subject contents, all students find Space and Figures definitely more difficult, but the gender-related difference is not significant within the Medium and the High latent classes. A gender gap in the student ability in favour of males is observed for Data and Predictions subject area, except within the Low class.

To conclude, at grade 10, the difference in mathematics performance by gender is significant from the mean ability levels, with a stable impact in the following levels. LCA groups confirm these results.

## 6. Discussion

The analyses clearly highlight, consistently with the literature results, the presence of a gender gap in performance in mathematics at all school grades, with boys ahead of girls. The use of a combination of different methods to analyse, more in depth, the achievement results along the entire distribution and to cluster the students into targeted groups with different levels of performance allowed to detail the impact of gender on the mathematics performance from different points of view adding new elements to the debate on the gender disparity.

With respect to the first research question about the difference between boys and girls both on a mean level and at different performance levels, the results confirm first that the difference on the overall mean Rasch score is significant al level 0.001 for all the grades. Moreover, the existence of gender-related differences on the mathematics performance for the several ability levels has been

verified by the use of quantile regression, showing that, at each grade, boys and girls performed differently along the ability distribution. These results have been confirmed by the latent class classification outcomes. In fact, the identification of three groups of students with different levels of performance enforces the hypothesis of a non-monotonic effect of gender along the performance distribution.

As far as the second research question is concerned, the analysis proved that gender gap is already present at primary school, evolving with some peculiarities through the next educational levels. At grade 5, at the end of primary school, the gap is already established, it is significant from low-medium performance levels remaining significant with a stable impact for the following ability levels. On the other hand, at grade 8, the gender gap is significant already for very low ability levels showing successively a stable impact from the medium levels. The results of grade 8 is consistent with international large-scale standardised tests showing that, in Italy, the gender gap didn't narrow like in other countries recently (Mullis et al., 2020). This finding needs further investigation to explore the possible impact of some gender stereotypes that could influence the emotional actions of girls in a delicate age such as adolescence. At grade 10, the difference in mathematics performance by gender is significant from the medium ability levels, remaining stable in the following levels. The comparisons made among the different school grades are descriptive, as proper longitudinal data are not available.

The third question about the pattern of performance by content domains was answered by investigating the gender gap within the latent classes. At grade 5, the differences in the performance are concentrated in the Medium class with Numbers and Data and Predictions domains in favour of boys, and Space and Figures in favour of girls. This last gap is also present in the High class. At grade 8, in the three classes the performances are not significantly different for the content domains, except for Space and Figures in favour of females, especially in the Low and the Medium classes.

For this reason, it should be preferable to support teaching/learning actions to improve the competence in all directions. Finally, at grade 10, the differences are more evident for the High class, with boys significantly outperforming girls in Numbers and Data and Predictions, and the Medium class with differences only in Data and Predictions. Within the Low class, the gap is significant for Space and Figures (in favour of males) and Relations and Functions (in favour of females) at the 5% level only. At this grade, the main differences highlighted can motivate activities to improve the learning of contents such as Numbers and Data and Predictions, especially for girls with medium and high levels of performance.

In general, the easiest subject at all levels for both genders and in all groups is Data and Predictions, whereas the hardest subject changes depending on the level and group. At primary school, it is Functions and Relations, while at lower secondary school, it is Numbers and, at upper secondary school, it is Spaces and Figures. The latter subject area merits particular consideration. In fact, the existing literature on gender studies frequently finds this to be an area in which boys score higher than girls (see, e.g., Halpern et al., 2007; Laurer et al., 2019), but the results of our analysis show the opposite instead. This aspect warrants further analysis. It could be due to how the items are formulated, although this aspect is not the main focus of this research.

## 7. Conclusion and future developments

The results can be used to develop targeted interventions along the entire performance scale and across the different school grades.

Low performances are associated with the lowest level of proficiency, meaning that students did not reach the minimum level of competences required at the end of the attended formative grade. Students with the highest level of performance demonstrate to acquire entirely the target of skills fixed for the specific grade. Medium performance characterised students that attained only

partially the needed competence. For each grade of school, females show different results, especially for medium and advanced acquired skills.

At the primary education stage, actions are required to reduce the gap at medium and higher-performing levels. At the end of lower secondary school, we can find significant differences already for the lowest level of performance, showing that girls struggle more to acquire basic concepts compared with boys. This result could affect girls' choices, pushing them towards less mathematics-oriented upper secondary schools. At this grade, it is fundamental to provide basic competences to help with future studies and to reduce gaps in each kind of school.

At grade 10, there is no significant gender gap among the students with the lowest level of competence acquired. The significant differences begin from the intermediate level of distribution. These results could be in part determined by behaviour, as previously described for lower school students. Girls struggle to narrow the attainment gap accumulated at previous study grades.

The novelty of the paper based on the analysis of content domains by student groups could be useful in the debate about teaching actions. The main objective of the study is to bring robust quantitative results allowing to make that decisions on possible educational actions which are competence of experts in mathematics teaching.

From a methodological point of view, some further developments are needed to confirm the results, above all in a longitudinal perspective, but also taking into account further factors such as cognitive learning domains and the impact of the question type. In particular, as the gender gap is different depending on both the grades and the content domains, an interesting development would be to study this topic in a longitudinal perspective.

In conclusion, our work offers room to enrich the debate around the international recommendations (see OECD, 2017b) to promote gender equality in education by ensuring equal

opportunities in making educational choices between boys and girls and making the study of STEM subjects equally inclusive and attractive.

## References


Adams, R. J., Wilson, M., & Wang, W.C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21, 1-23.

Aeschlimann, B., Herzog, W., & Makarova, E. (2016). How to foster students' motivation in mathematics and science classes and promote students' STEM career choice. A study in Swiss high schools. *International Journal of Educational Research*, 79, 31-41.

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.

AlmaLaurea, (2020). XXII *Indagine - Profilo dei Laureati 2019*, *Rapporto 2020*.
https://www.almalaurea.it/universita/profilo/profilo2019/volume

Arnup, J. L., Murrihy, C., Roodenburg, J., & McLean, L. A. (2013). Cognitive style and gender differences in children's mathematics achievement. *Educational Studies*, 39(3), 355-368.Battista, M. T. (1990). Spatial visualisation and gender differences in high school geometry. *Journal for Research in Mathematics Education,* 21(1), 47-60.

Baye, A., & Monseur, C. (2016). Gender differences in variability and extreme scores in an international context. *Large-scale Assessments in Education*, 4(1).

Bedard, K., & Cho, I. (2010). Early gender test score gaps across OECD countries. *Economics of Education Review*, 29(3), 348-363.

Byrnes, J. P. (2005). Gender differences in math: Cognitive processes in an expanded framework. In A. M. Gallagher & J. C. Kaufman (Eds.), *Gender differences in mathematics. An integrative psychological approach* (pp. 73-98). New York: Cambridge University Press.

Cascella, C. (2015). Male and female performance in mathematics: Empirical evidence from Italy. *International Journal of Interdisciplinary Educational Studies,* 9(3-4), 1-9.

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, Article 100819.

Cleary, T. A. (1992). Gender differences in aptitude and achievement test scores. In Sex equity in educational opportunity, achievement, and testing: Proceedings of the 19.91 ETS Invitational Conference (pp. 51-90). Princeton, NJ: Educational Testing Service.

Cornwell, C., Mustard, D., & Van Pairs, J. (2013). Gender disparities in test scores and teacher assessments: Evidence from primary school. *Journal of Human Resources*, 48(1), 236-264.

Contini, D., Di Tommaso, M. L., & Mendolia, S. (2017). The gender gap in mathematics achievement: Evidence from Italian data. *Economics of Education Review*, 58, 32–42.

Costanzo, A., & Desimoni, M. (2017). Beyond the mean estimate: a quantile regression analysis of inequalities in educational outcomes using INVALSI survey data. *Large-Scale Assessments in Education*, 5(14), 1-25.

Davino, F., Furno, M., Vistocco, D. (2013). *Quantile Regression: Theory and Applications*, Hoboken, NJ: John Wiley & Sons.

Dayton, C.M., & Macready, G. B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association*, 83, 173–178.

DeMars, C. E. (1998).Gender differences in mathematics and science on a high school proficiency exam: The role of response format. *Applied Measurement in Education*, *11*(3), 279-299.

De Simone, G. (2013). Render unto primary the things which are primary's: Inherited and fresh learning divides in Italian lower secondary education. *Economics of Education Review*, 35, 12-23

Erdoğan, A., Baloğlu, M., & Kesici, Ş. (2011). Gender differences in geometry and mathematics achievement and self-efficacy beliefs in geometry. *Eurasian Journal of Educational Research*, 43, 188-205.

Falorsi, P., Falzetti, P., & Ricci, R. (2019). *Le metodologie di campionamento e scomposizione della devianza nelle rilevazioni nazionali dell'INVALSI* (Methods of sampling and decomposition of the deviance in INVALSI national surveys). Milano: Franco Angeli.

Feingold, A. (1992). Sex differences in variability in intellectual abilities: A new look at an old controversy. *Review of Educational Research*, 62, 61–84.

Fennema, E. (1974). Mathematics learning and the sexes: A review. *Journal for Research in Mathematics Education*, 5, 126-139.

Fennema, E., & Sherman, J. (1977). Sex-related differences in mathematics achievement, spatial visualization and affective factors. *American Educational Research Journal, 14*, 51-71.

Fennema, E., & Sherman, J. (1978). Sex-related differences in mathematics achievement and related factors: A further study. *Journal for Research in Mathematics Education, 9*(3), 189-203.

Ferretti, F., & Giberti, C. (2020). The properties of powers: Didactic contract and gender gap. *International Journal of Science and Mathematics Education*, in press, DOI: 10.1007/s10763-020-10130-5.

Fryer, R. G., & Levitt, S. D. (2010). An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics,* 2(2), 210-240.

Gallagher, A. M. (1998). Gender and antecedents of performance in mathematics testing. *Teachers College Record*, 100, 29.

Gallagher, A. M., & Kaufman, J. C. (2005). Gender differences in mathematics. What we know and what we need to know. In A. M. Gallagher & J. C. Kaufman (Eds.), *Gender differences in mathematics. An integrative psychological approach* (pp. 316-332). New York: Cambridge University Press.

Geary, D. C. (1996). Sexual selection and sex differences in mathematical abilities. *Behavioral and Brain Sciences*, 19, 229-284.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61 (2), 215–231.

Hadjar, A., Krolak-Schwerdt, S., Priem K., & Glock, S. (2014). Gender and educational achievement. *Educational Research*, 56(2), 117-125.

Hagenaars, J. A., & McCutcheon, A. L. (2002). *Applied Latent Class Analysis*. Cambridge: Cambridge University Press.

Halpern, D. E (1997). Sex differences in intelligence: Implications for education. *American Psychologist*, 52, 1091-11

Halpern, D. F. (1986). Sex Differences in Cognitive Abilities (1st ed.). Hillsdale, NJ: Erlbaum.

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The Science of sex difference in science and mathematics. *Psychological Science in the Public Interest*, 8(1), 1–51.

Heckman, J. J., & Kautz, T. (2012). Hard evidence on soft skills. *Labour Economics,* 19(4), 451–464.

Heckman, J. J., & Kautz, T. (2014). Achievement tests and the role of character in American life. In J. J. Heckman, J. E. Humphries, & T. Kautz (Eds.), *The Myth of Achievement Tests: The GED and the Role of Character in American Life* (pp. 3-56). Chicago: University of Chicago Press.

Hedges, L. V., & Friedman, L. (1993). Gender differences in variability in intellectual abilities: A reanalysis of Feingold's results. *Review of Educational Research, 63*(1), 94–105.

Hyde, J. S., Fennema, E., & Lamon, S. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bullettin*, 107(2), 139-155.

INVALSI (2017). Rilevazioni Nazionali degli Apprendimenti 2016-2017- Rapporto Risultati. https://invalsi-areaprove.cineca.it/docs/file/Rapporto_Prove_INVALSI_2017.pdf

INVALSI (2018). Rapporto Prove INVALSI 2018. Rapporto Nazionale. https://invalsi-areaprove.cineca.it/docs/2018/Rapporto_prove_INVALSI_2018.pdf

INVALSI (2019a). Rapporto Prove INVALSI 2019. Rapporto Nazionale. https://invalsi-areaprove.cineca.it/docs/2019/Rapporto_prove_INVALSI_2019.pdf

INVALSI (2019b). OCSE PISA 2018. I Risultati degli Studenti Italiani in Lettura, Matematica e Scienze. Rapporto Nazionale. https://www.invalsi.it/invalsi/ri/pisa2018/docris/2019/Rapporto_Nazionale.pdf

Karakolidis, A., Vasiliki, P., & Emvalotis, A. (2016). Examining students' achievement in mathematics: A multilevel analysis of the Programme for International Student Assessment (PISA) 2012 data for Greece. *International Journal of Educational Research*, 79, 106-115.

Koenker, R. (2010), *Quantile Regression*, Cambridge: Cambridge University Press.

Koenker, R. (2019). Quantreg: Quantile regression R package version 5.54 https://cran.r-project.org/web/packages/quantreg/

Koenker, R. W., & Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1), 33-50.

Lauer, J. E., Yhang, E., & Lourenco, S. F. (2019). The development of gender differences in spatial reasoning: A meta-analytic review. *Psychological Bulletin*, 145(6), 537–565.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent Structure Analysis*. Boston: Houghton Mifflin.

Linzer, D. A., & Lewis, J. B. (2011). poLCA: An R Package for Polytomous Variable Latent Class Analysis. *Journal of Statistical Software,* 42 (10), 2-29.

Matteucci, M., & Mignani, S. (2011). Gender differences in performance in mathematics at the end of lower secondary school in Italy. *Learning and Individual Differences*, 21, 543-548.

Meinck, S., & Brese, F. (2019). Trends in gender gaps: Using 20 years of evidence from TIMSS. *Large-scale Assessments in Education,* 7(8), 1-23.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 international results in mathematics*. Retrieved from Boston College, 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.

Muzzatti, B., & Agnoli, F. (2007). Gender and mathematics: Attitudes and stereotype threat susceptibility in Italian children. *Developmental Psychology*, 43(3), 747-759.

Nemeth, B. (2007). Measurement of the development of spatial ability by the Mental Cutting Test. *Annales Mathematicae et Informaticae,* 34, 123-128.

Neuschmidt, O., Barth, J., & Hastedt, D. (2008). Trends in gender differences in mathematics and science (TIMSS 1995–2003). *Studies in Educational Evaluation*, 34(2), 56-72.

Nguyen, H.-H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta analysis of experimental evidence. *Journal of Applied Psychology*, 93, 1314-1334.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

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

OECD (2017a). *Education at a Glance 2017: OECD Indicators*. Paris: OECD Publishing. http://dx.doi.org/10.1787/eag-2017-en

OECD (2017b), *2013 OECD Recommendation of the Council on Gender Equality in Education, Employment and Entrepreneurship*, OECD Publishing, Paris.

OECD (2019a). *Education at a Glance 2019: OECD Indicators*. Paris: OECD Publishing https://doi.org/10.1787/f8d7880d-en

OECD (2019b). PISA 2018 Results (Volume II): Where All Students Can Succeed, Paris: OECD Publishing https://doi.org/10.1787/b5fd1b8f-en

Peterson, P., & Fennema, E. (1985). Effective teaching, students engagement in classroom activities, and sex-related differences in learning mathematics. American Educational Research Journal, 22(63), 309-335.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research.

R Core Team (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. URL http://www.R-project.org/

Robinson-Cimpian, J. P., Lubienski, S. T., Ganley, C. M., & Copur-Gencturk, Y. (2014). Teachers' perceptions of students' mathematics proficiency may exacerbate early gender gaps in achievement. *Developmental Psychology*, 50(4), 1262-1281.

Robinson, J. P., & Lubienski, S. T (2011). The development of gender achievement gaps in mathematics and reading during elementary and middle school: Examining direct cognitive assessments and teacher ratings. *American Educational Research Journal,* 48(2), 268-302.

Sansone, D. (2018). Teacher characteristics, student beliefs, and the gender gap in STEM fields. *Educational Evaluation and Policy Analysis*, 41(2), 127-144.

Strand, S., Deary, I. J., & Smith, P. (2006). Sex differences in cognitive ability test scores: A UK national picture. *British Journal of Educational Psychology*, 76(3), 463–480.

UNESCO (2017). Cracking the code: Girls' and Women's Education in Science, Technology, Engineering and Mathematics (STEM). https://unesdoc.unesco.org/image s/0025/00253 4/25347 9E.pdf.

Willingham, W.W., & Cole, N.S. (1997). Gender and Fair Assessment. New York: Routledge.