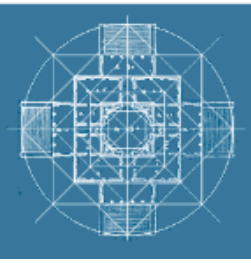


ITALIAN JOURNAL OF SOCIOLOGY OF EDUCATION

Editor-in-Chief: Silvio Scanagatta | ISSN 2035-4983

The North-South Divide in School Grading Standards: New Evidence from National Assessments of the Italian Student Population

Gianluca Argentin[*] *and Moris Triventi*[**]

Authors' information

[*]Department of Sociology, Catholic University of the Sacred Heart, Italy.
[**]Department of Political and Social Sciences, European University Institute, Italy.

Contact authors' email addresses

[*]gianluca.argentin@unicatt.it
[**]moris.triventi@eui.eu

Article first published online

June 2015

HOW TO CITE

Argentin, G., & Triventi, M. (2015). The North-South Divide in School Grading Standards: New Evidence from National Assessments of the Italian Student Population. *Italian Journal of Sociology of Education*, *7*(2), 157-185. Retrieved from http://journals.padovauniversitypress.it/ijse/content/north-south-divide-school-grading-standards-new-evidence-national-assessments-italian

PADOVA UNIVERSITY PRESS

The North-South Divide in School Grading Standards: New Evidence from National Assessments of the Italian Student Population

Gianluca Argentin[*] *and Moris Triventi*[**]

Abstract: Even if marks are crucial for students' educational careers and school-related decisions and although grading standards are a relevant topic in public debate about Italian education, in our country this research topic has not attracted much attention. In this article we investigate heterogeneity across Italian macro-regions in grading standards (degree of strictness in attributing marks by teachers) and in the coherence between teachers' marks and students' test scores. We use data from INVALSI-SNV on the whole student population in the 5th, 6th and 10th grade in 2011/12, with relevant information on two subjects (Italian and mathematics). We detect that Southern regions are characterized by what seems higher generosity in grading students, who display lower performance in the INVALSI assessment compared to their counterparts with the same marks and socio-demographic profile. Moreover, this generosity in attributing marks seems stronger for higher marks (9 and 10) and in mathematics, especially in lower secondary schools and lyceums. At the same time, our analysis underlines that the North-South divide is crucial but provides only a partial view of the phenomenon: indeed, we find striking differences in grading standards among Italian provinces even within macro-regions. We discuss the main implications of such geographical heterogeneity for the Italian educational system.

Keywords: grading standards, teachers' marks, test scores, geographical inequalities

[*]Department of Sociology, Catholic University of the Sacred Heart, Italy. E-mail: gianluca.argentin@unicatt.it

[**]Department of Political and Social Sciences, European University Institute, Italy. E-mail: moris.triventi@eui.eu

Introduction

Grading is an educational practice that affects the lives of millions of students in various school levels and it is a crucial aspect of the relationship with their teachers. In education, teachers' marks can perform many functions, but one of the most important is to provide students feedback about their learning and level of knowledge in a particular discipline. Teachers' marks can be considered important from a sociological point of view because they are a visible feedback for students and their families, who cannot easily and directly observe educational knowledge, skills and achievement.

Marks are powerful signals and they may influence students' motivation to study, self-confidence, and effort. Moreover, marks provide information that can contribute to shape the educational choices made by students and their parents (OECD, 2012; Gasperoni, 1998). Final marks matter also outside of school: they represent one of the criteria taken into account by universities and fields of study with restrictions at access and they can be used by employers as a signal about the skills and quality of a potential candidate for a job (Johnes, 2004).

There is evidence that teachers' grading practices are the result of a multifaceted assessment process, which does not only reflect the objective level of skills achieved by students, but also the perceived students' effort, motivation, and even their behaviour (OECD, 2012; 2013). Marks reflect daily interactions between teachers and students and can be used by these actors as incentives in a "strategic" way (Costrell, 1994). Moreover marks could be affected by teachers' and students' social group identities and may contribute to the reproduction of social inequalities (Bowles and Gintis, 1976).

Nonetheless, it is important to examine to what extent teachers' evaluations reflect their students' skills. The relevance of this issue for educational policies has also been emphasized in the report "PISA in Focus' number 26" on Grade expectations, where the OECD (2013) examines the misalignment between students' performance in the test and their reported marks. Moreover, several articles focused on the US case, suggests that grading standards can have non-negligible effects on the pupils' subsequent learning and skills development (Betts and Grogger, 2003; Figlio and Lucas, 2004).

A well-established way to investigate teachers' grading standards is to compare their marks with student achievement in standardized tests administered by independent institutional actors assessing the same subjects. Although school grades have lower costs and can be attributed frequently, standardized tests are designed to detect students' competencies in the most neutral and "objective" way as possible and therefore may be considered as a reference measure and a good proxy of pupils' school-related skills. This external reference can be helpful in establishing the severity of teachers' grading and their fairness in ranking students along the grades distribution (Bonesrønning, 1999; Lindahl, 2007).

The literature on grading standards is mainly based on the comparison between marks, a multi-purpose multi-dimensional measure, and test scores, a flat mono-dimensional and subject specific measure[1] of students' knowledge. The comparison is quite problematic and the authors analysing this research field are aware of this. Nonetheless, it is possible to take into account the intrinsic difference between marks and scores in our interpretation of results. We could reflect upon results coming from marks-scores comparison not only in terms of grading standards, but also in terms of the signalling value linked to school marks.

We mean that the distance between marks and scores could be interpreted not only in terms of teachers' severity/generosity and coherence, but also in terms of which contents are actually signalled by marks: to what extent do they really signal student performance? How much do marks signal something else (i.e. their behaviour in school, their effort, their relative performance compared to classmates, etc.)? Summing up, comparing students' marks and scores, we are in the condition to reflect about marks as a signal: both in terms of teachers' grading standards and in terms of marks' meaning and content.

While there is a body of research in other countries that analysed the variability and the effects of schools grading standards (among others Bonesrønning, 1999; 2004; 2008; Betts & Grogger, 2003; Lavy, 2008; Lindahl, 2007; Dardanoni, Modica & Pennisi, 2012), in our country empirical evidence

[1] Clearly also this measure is less monolithic than it seems and the error sources are many (Koretz, 2008).

on this topic is less developed, partly because of the late introduction of learning assessment through standardized tests at national level, and partly because of the Italian educational scientists community not being used to analysing data coming from students' standardized assessments.

The main goal of this article is to provide empirical evidence on the relationship between teachers' school marks and students' performance in the standardized tests administered by the SNV-INVALSI (National Evaluation of the School System). Compared to previous research which focuses mostly on local samples of students, we provide (according to our knowledge) the first large scale and systematic assessment of teachers' grading standards, at the same time in two different subjects (reading/Italian and mathematics) and across three different educational levels (primary, lower secondary, and upper secondary education), examining data on about one million of students. More specifically, we are interested in investigating whether and to what extent grading standards and marks contents are homogenous across geographical areas, with a focus on the North/Centre versus South divide[2].

This geographical cleavage is a key feature of the Italian educational system (Bratti, Checchi, & Filippin, 2007; INVALSI, 2010) and already emerged as relevant also for the grading topic (see next section). While existing studies suggest variations across macro-geographical areas in the degree of strictness in teachers' standards, it seems that the major distance is found between the Northern and Central regions, on one side, and the Southern and Islands regions, on the other. Overall, we are interested in testing whether:

a. students attending school in Southern regions are assessed differently than in Northern regions, namely if the average level of scores associated to students' with the same marks is homogeneous across geographical area;

b. the coherence between teachers' assessment and students' performance (measured via standardized test) differs between North/Centre and South (the correlation between marks and scores in the two areas). In order to check to

[2] All the analyses are based on a comparison between North and Centre, collapsed in a unique macro-region, and the South and the Islands that constitute the second macro-area. For simplicity, we refer sometimes to them using the terms "North" and "South" or "Northern regions" and "Southern regions".

what extent these macro-geographical areas are internally homogeneous, we also perform, for the first time in Italy, a disaggregated analysis at the province level.

The issue of diverging grading standards among macro-regions is particularly relevant for several reasons. First of all, as stated above, one of the functions of grades is to provide information or a signal to pupils and their families about the student's competencies. If families in some parts of the country are receiving a distorted signal, because marks are too generous/severe or do not reflect subject knowledge but something else, this is likely to end in a partly inefficient allocation of students on the basis of their abilities and competencies. Knowing whether this source of misallocation in the educational investment differs among macro-regions is relevant in terms of public policies. Second, this topic matters for equity among students coming from different geographical areas. For instance, the final mark in upper secondary education is used as a criterion for admission in university programmes with access restrictions. Heterogeneous grading standards across geographical areas will imply different opportunities between students who attended school in different regions, ending up with different marks despite the same level of competencies (or *vice versa*).

Finally, the topic of grading standards is relevant for the recent debate about the need for external committee members or the use of a standardized assessment in the final national exams in upper secondary education. The article is organized as follows: in the next section we review the empirical literature on grading standards and related teachers' assessment practices in Italy. In the third section we discuss the research aims of the analyses and we outline some general hypotheses. The fourth section presents the data, variables and methods used in the analysis. The fifth section presents descriptive statistics and results from multivariate models. Finally, we derive some conclusions and implications for research and policies coming from our analyses.

Previous studies in Italy

A few years ago Bolletta (2001) designed and conducted an extensive empirical survey on State examinations at the end of upper secondary education. The study adopted a docimological approach aiming at identifying whether different teachers evaluate students' exams in similar or heterogeneous ways. Among many results, one striking finding refers to the mathematics assessment, where eleven teachers have been asked to independently grade 20 exams by high school leavers. This work found a large variation in the minimum and maximum mark attributed by teachers to the same test, a range amounting to 7 points on a total scale of 15 points.

A similar finding was found by a recent survey on the "Italian" written examination of the upper secondary education final exams in 2007 (INVALSI and Accademia della Crusca, 2009), in which the exams were evaluated by the Examination Board, two additional independent teachers (free evaluators) and, in a limited number of cases, by a further teacher operating on the basis of precise indications (evaluator with correction card). The report showed a considerable variation in the evaluation standards adopted by the different types of judgers, with only a 23% of agreement between the grades of these different categories of teachers and a systematic lower severity by the Examination board compared to the "free evaluators". A local study, conducted in Trentino in 1999, focused once again on State examinations at the end of upper secondary education. In this case, the analyses showed that the teachers grading standards were more relaxed in the oral exam for students who, after the written tests, were at the threshold between passing or failing the exam (Argentin and Tamanini. 2001).

A second approach in studying grading standards, has been adopted by Sestito and Tonello (2011), who analysed the relationship between upper secondary final marks and performance in the standardize tests to enter the faculty of Medicine at university. Given that there is a unique national test to enter Medicine, it is administered in the same moment across the institutions and the test content includes "general knowledge items", the existence of a low correlation between the two measures can be regarded as indirect evidence of heterogeneous grading standards across different schools and geographical

areas. In 2009 the correlation was found to be relatively small, amounting to 0.32, which suggests that the diploma final mark is not a reliable measure of the academic preparation of students leaving upper secondary education, at least among those who intend to study Medicine. Another paper by Checchi and colleagues (2011) looked at the relationship between the specific high school where students obtained their diploma and the subsequent performance in higher education. This study was conducted only in Lombardy, one of the largest and most developed Northern regions. Even if the degree of heterogeneity in this area is relatively small, there is evidence that the effect of final mark on subsequent results at university, in terms of credit accumulation and exam marks, is not homogeneous across schools, even within the same educational track. Hence, marks attributed from different schools could be stronger or weaker predictors of academic performance even within the same region. In the same regional context, Iacus and Porro (2011) investigated the grading practices of teachers in lower secondary schools, finding that not only the overall level of severity varies across schools but also the "style" of attributing marks is differentiate across classes/teachers.

Finally, some studies analysed various editions of the PISA (Programme for International Student Assessment) survey, comparing the average grade obtained in the annual report card preceding PISA assessment and the score achieved by students in standardized tests. IReR (2006; 2008) reported a positive relationship between teacher's marks and mathematics skills, but also a great diversity of grading standards by type of school and region. The assessment standards are more stringent in the lyceums and in schools located in Northern Italy, while they are less severe in vocational schools and in Southern Italy. Similar results were found by Gay and Triventi (2011), who analysed the variability of grading standards in reading using data from PISA 2009. They found that in the evaluation, standards become less stringent moving from the Northern regions and Centre to the South and Islands. In particular, in Southern regions teachers tend to attribute the "pass" mark more easily than teachers in the Northern regions, and this result does not change accounting for the possibly different distributions of pupils' individual characteristics and types of school.

Summing up, pre-existing studies showed that there is high variety in grading standards among teachers, schools and regions in Italy, as in several other countries (Dardanoni, Modica & Pennisi, 2012). Therefore investigating this topic once again seems relevant for the Italian educational research. The new population datasets made available by INVALSI are a great opportunity to make advancement in our knowledge of this field.

Research aims

The main aim of this article is to examine in detail whether and how teachers' grading standards and marks signal content vary across geographical areas in the Italian educational system. The main focus is on the differences between two macro-geographical areas: the Northern and Southern regions. As shown in the previous section, indeed, existing research seems to suggest the existence of differentiated grading practices between these areas, with the South displaying lower overall achievement in standardized tests, but higher levels of marks given by teachers. Nonetheless pre-existing evidence is based only on upper secondary schools. Furthermore, we also examine a more fine-grained disarticulation of geographical areas, analysing the variation in grading standards at the province level.

Compared to existing studies, we focus on a different data source, namely the INVALSI-SNV (*Sistema Nazionale di Valutazione*) on the entire national student population assessed in three school grades in primary, lower and upper secondary education. Thanks to this larger and richer dataset, we are able to expand existing knowledge on grading standards investigating whether the differences between the Northern and Southern regions are stable or differs across school levels. Within upper secondary level, we also study grading standards across school tracks, which constitute widely different environments in the Italian educational systems (in terms of curriculum, requirements, teachers' characteristics) (Gambetta, 1987; Gasperoni, 1998; Panichella and Triventi, 2014) and have also been found to be an important source of heterogeneity in students' competencies (Checchi and Flabbi, 2007; Bratti, Checchi, & Filippin, 2007). Furthermore, we are in the position to

simultaneously analyse the same student population in two different subjects, Italian-reading and mathematics, previously investigated only in separate studies and among students of different age, school grade and region. Hence we can assess whether the grading process differs systematically among teachers of Italian-reading and Mathematics (the two most important subjects in the Italian school system). One advantage of the INVALSI data over the PISA data is that in the former information on students' mark assigned by teachers is derived from administrative sources, while in the latter is reported by the student him/herself (leading to a probably larger measurement error and potential unobservable systematic distortions).

Finally, our analyses investigate two different dimensions in teachers' grading practices: a) *the grading standards*, which are intended to measure how much the teachers are severe/rigorous when attributing marks to their students and how much marks really refer to students skills or to other unobserved students' characteristics; b) *the coherence* between teachers' marks and students' competencies, which is informative about the relationship between teachers ranking and students position in the INVALSI score distribution. More information on how to empirically measure these two concepts is provided in the next section.

Data, variables and methods

As anticipated, we use data from the National Assessment Program INVALSI-SNV (*Sistema Nazionale di Valutazione*) for the 2011-12 school year[3]. The data for each school level considered in our research are provided in three different data sources, each containing three datasets: 1) test scores in reading; 2) test scores in mathematics; 3) student's questionnaire. Files 1 and 2 report information on the detailed answers of the students to the single items constituting the INVALSI test, various measures which summarize the total

[3] The data are a special release provided by INVALSI to the authors as winners of the "Concorso Idee 2013" with the project entitled "Come mi giudichi? Analisi delle pratiche e degli standard di attribuzione dei voti agli studenti nelle scuole italiane".

student score and additional information on the individuals provided by schools on the basis of their administrative registers. File 3 reports the answers by students to an *ad hoc* questionnaire about their social background, school environment, their behaviour and attitudes. INVALSI tests and questionnaire for the 2011-12 school year were administered the 9th and 11th May 2012 in the fifth grade (last year of primary school), the 10th May 2012 in the sixth grade (first year of lower secondary school), and the 16th May in the tenth grade (second year of upper secondary school). The expected time for filling each test was differentiated according to the school level but uniform for all students within the same school level. The administration of the test and questionnaire was conducted by a school teacher, usually not teaching the tested subject in the surveyed class. In a random sample of targeted schools the administration took place in the presence of an outside observer. Thanks to this external control it was possible to use the random sample data to correct the national assessment for bias due to cheating, according to the INVALSI procedure (INVALSI 2012).

The total number of classes surveyed by the SNV amounts to 29,804 in V primary, 27,402 in II lower secondary, 24,751 in II upper secondary. The total number of students in these three school grades is respectively 558,371, 611,663, and 533,260.

In order to conduct our analyses we constructed a pooled dataset of students assessed in the three grades. We have also restricted the analytical sample in the following way. First, for descriptive analyses we considered cases with valid information on students' test scores and teachers' marks both in Italian and mathematics. For the multivariate analyses we further restricted the sample adopting a list-wise deletion of cases with missing information on the variables used as covariates in our regression models (see below). The analytical sample size includes 308,334 cases in the 5th grade, 321,892 in the 6th grade, and 279,524 in the 10th grade, for a total of 909,750 cases.

The four key variables considered in our study are the scores obtained by students in the INVALSI standardized tests of Italian and Mathematics and teachers' marks in the same subjects. Students' performance is measured by INVALSI applying IRT Rasch model to their answers to the SNV tests; they are standardized to have a mean of 200 and a standard deviation of 40 in the

whole original sample. We use the scores adjusted for potential cheating, provided in the datasets. To measure the teachers' evaluation of the same students, we rely on marks in mid-term reporting card (February 2012) in the same two subjects (Italian and Mathematics).[4] The original variables range from 1 to 10. In the analyses, we restrict the estimation on students who receive a mark between 3 and 10, to exclude outliers and values that could probably be due to data entry mistakes.[5]

The form used by INVALSI to collect mid-term marks from schools asked both the "written marks" and "oral marks", allowing (if existing) separate grading on the basis of different assessment tools. We preferred the "written marks", due to the fact that also the test is administered in a paper and pencil setting. For students in the fifth and sixth grade with missing "written marks", we used the information on the "oral marks", since the correlation between the two types of marks (for those who have both) amounts to around .97. This is probably due to the fact that it is quite unusual to have separate marks for written and oral assessments in primary and lower secondary schools; it is plausible that, in many cases, schools simply reported twice the same mark. We did not implement this substitution among students in the 10th grade, because in this case the correlation is only .68 and in upper secondary schools it is more common to have separate written and oral marks.

To identify geographical areas we used a dummy variable distinguishing the North/Centre from the South/Isles. The classification of Italian regions into these two macro-areas is the standard proposed by ISTAT. For additional analysis, we use the province in which the school is located as a fine-grained indicator of geographical area[6].

We included several control variables in the multivariate analyses. They are: gender, migration status (natives, first generation, second generation), highest education level attained by student's parents (university, upper secondary, lower secondary education or less), highest social class attained by parents

[4] Unfortunately, the datasets do not provide information on the final year marks.
[5] However, the proportion of cases excluded from the analysis is less than 0.2% and this choice does not substantially affect our results.
[6] The classification corresponds to the one reported in the 2001 Census of the Italian population.

(high bourgeoisie, entrepreneurs, white collars, autonomous workers, working class, unemployed), number of books at home (no or very few, one shelf, one bookcase, two bookcases, three or more bookcases), area of birth and whether the school was included in the random sample with an outside observer (being this a factor influencing students' performance).

Regarding methods, we developed several indicators to measure the different grading practices in the various geographical areas of the country. First of all, it is important to distinguish two dimensions in the grading practices: the first one is the *coherence* between teachers' marks awarded to their students and the performance obtained in standardized test by the same students. The coherence indicates the degree by which the evaluations based on standardized assessments and those based on the summary evaluation by teachers are correlated, coinciding or deviating. The second dimension instead refers to the *grading standards* adopted by the teachers, which are conceived as the absolute score obtained by students who have obtained the same mark. The logic is the following: if two students with the same mark in a given subject obtain different scores, the one with the lower score has been exposed to more generous standards or is assessed by a teacher basing his/her marks less on student knowledge and more on other contents (such as students effort, self-control, etc.).

In this work we employed the linear Pearson's correlation between between test scores and teachers' marks to measure their degree of agreement. If the two variables co-vary in the same direction, we should find a coefficient ranging from 0 to 1. The larger the value is, the higher the agreement is between the two tools for student evaluations (teacher mark and INVALSI test).

Looking at the second dimension, we used two indicators. The first measure is derived from an OLS linear regression model in which the dependent variable is the score obtained by the student in the standardized test ($SCORE_i$), which is expressed as a function of the mark obtained in the report card ($MARK_i$), the dummy variables that indicate the geographical area ($AREA_i$, which is the macro-region in the first analysis and the province in the second), a vector of control variables described in the previous section (Z_i) and the residuals (ε_i).

$$E(SCORE_i) = \alpha + \sum_{k=1}^{K-1} \beta_k AREA_{ik} + \gamma\, MARK_i + \sum_{k=1}^{K-1} \partial_k Z_i + \varepsilon_i$$

This model provides a measure of the differences among geographical areas in the degree of adherence of teachers' grading to students' subject knowledge, captured by the regression coefficients β_k. These coefficients are associated with macro-areas in the first analysis (K=2) and with provinces in the second analysis (K=103). The coefficients indicate the average difference in the predicted score obtained in the standardized assessment across areas, controlling for teachers' marks and other individual-level variables. When analysing the North-South divide, a positive coefficient indicates that in the South there are stricter standards compared to the North, while a negative coefficient means that teachers in the South adopt more generous standards or they give less importance to students' knowledge and more to other aspects when attributing marks.

For measuring the North-South divide in grading standards, we adopt the standard parametrization, reporting the difference between the expected score obtained by students in the South and those in the North (the omitted reference category). When analysing differences across provinces, we will use a different parametrization, which produces results easier to present and discuss, given the large number of categories. We compute and report the predicted difference in grading standards between each Italian province and the national average (effect-coding). In this way, we can obtain an estimate of the level of severity of the evaluation standards for each province, without the need to omit an arbitrary reference province.

The first indicator assumes that the differences between geographical areas are homogeneous along the marks distribution, but this is likely not to be the case, as discussed by Iacus and Porro (2011). For instance, teachers in the South can be more generous when attributing high marks, while they could be more rigid when giving lower marks. In order to test whether this is the case, we estimate the following regression model:

$$E(SCORE_i) = \alpha + \sum_{k=1}^{K-1} \delta_k (SOUTH_i \times MARK_i) + \beta SOUTH_i \sum_{k=1}^{K-1} \gamma_k MARK_i + \sum_{k=1}^{K-1} \partial_k \mathbf{Z}_i + \varepsilon_i$$

where we add an interaction between the geographical area (*SOUTH*) and the dummy indicators for the different marks (*MARKS*), included as a categorical variable. In this case the parameters of interest are the δ_k, which indicate the average difference in the predicted test score between students in the South and in the North according to the mark obtained in the mid-term report. Given the high number of estimated parameters, we adopt this method only to examine the North-South differences and not the differences between provinces.

Results

In this section we present the main results of the empirical analysis. We will begin with some comments on the basic descriptive statistics on the INVALSI data. Then, we will investigate the coherence between grades and scores across macro-regions. Then, we will discuss the results regarding the grading standards across macro-regions and provinces in our educational system[7].

Descriptive statistics: teachers' marks and test scores

First of all, it is interesting to inspect the overall distribution of teachers' marks. There is a clear difference in the overall level of teachers' marks according to the school level: the average mark is higher in the fifth grade (7.6 and 7.7 respectively for Italian and mathematics), while it is substantially smaller in the sixth grade, reaching 6.6 and 6.7 in Italian and mathematics. Then, it falls further in the second year of upper secondary, where the average grade in the mid-term report card is slightly below the "pass" threshold in both subjects (5.9 for Italian and 5.7 for mathematics). These averages are likely to

[7] For parsimony and to ease the detection of the main patterns of interest the results are presented in graphical form. The complete results in tabular form are available from the authors upon request.

be smaller than marks assigned at the end of the scholastic year, since – among teachers – it is common practice to be stricter in grading in the first rather than in the last school report. Looking at the tracks in upper secondary education, teachers' marks are higher in lyceums (6.1 and 5.9), while lower in technical institutes (5.8 and 5.5) and in vocational schools (5.7 and 5.5).

This does not mean necessarily that teachers in technical/vocational schools are stricter in attributing marks, given that they could teach to lower-ability students. These figures also suggest that, on average, teachers' marks in mathematics are lower compared to those in Italian, with the exception of primary education. Besides exploring the average grade, it can be useful to describe the percentage distribution of marks in Italian and mathematics according to the school level. Figure 1 shows that these distributions have the form of a bell centred on a single or two contiguous modal values.

The distributions appear gradually slipped to the left, going from lower to higher school levels. The distribution in the fifth grade is shifted upwards compared to the others: the most frequent mark is 8, the "3" is mostly absent and only very few students have obtained a "4" in the report card. Furthermore, between 25% and 28% of students – depending on the subject – has achieved an excellent evaluation, at least equal to 9. The distribution of marks in the first year of lower secondary education is instead shifted slightly to the left: the modal value is the "7", followed by "6" (pass).

Also in this case the "3" are very few, while there is a non-negligible proportion with the "4", amounting to 3.2% in Italian and 5.6% in mathematics. The share of students with at least 9 in the report card is much lower than in elementary school: it reaches 5% in Italian and does not exceed 10% in mathematics.

Finally, in the 10th grade the distribution is shifted further down: the modal vote in Italian is, in fact, the "6", which contains as much as 40% of students. In mathematics around 25% received 6 and about 23% got 5 in the report card. The percentage of students with excellent grades (at least 9) is very small at this school level: less than 1% in Italian and around 3.5% in mathematics.

Figure 1. Average test scores in Italian and in Mathematics according to teachers' marks

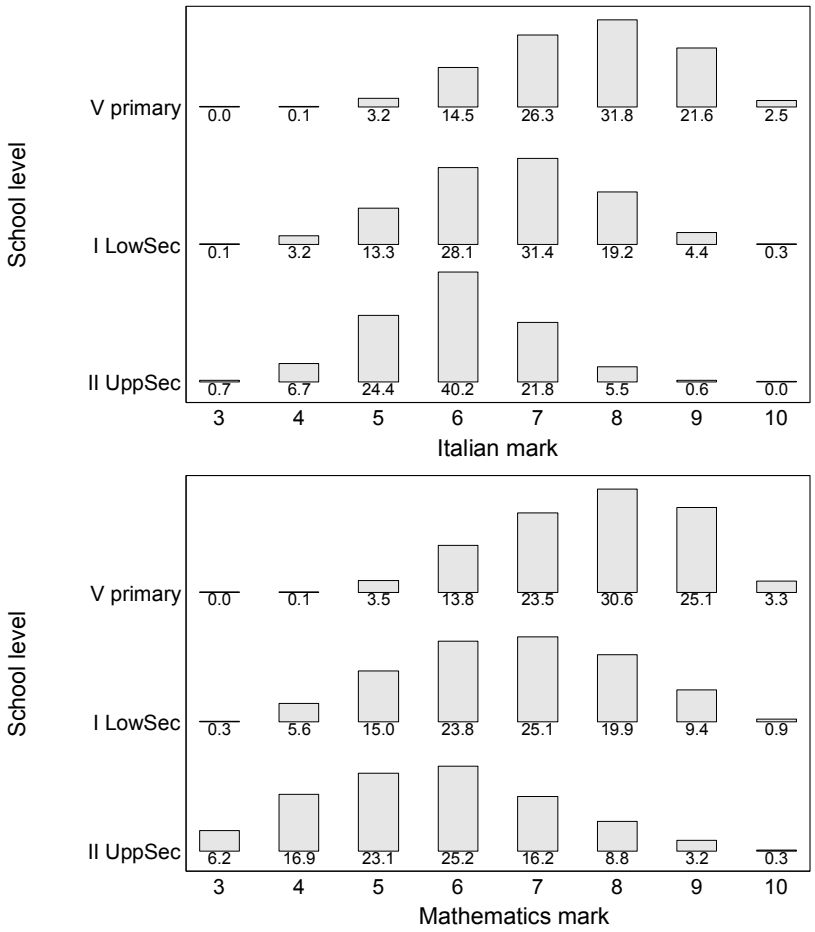
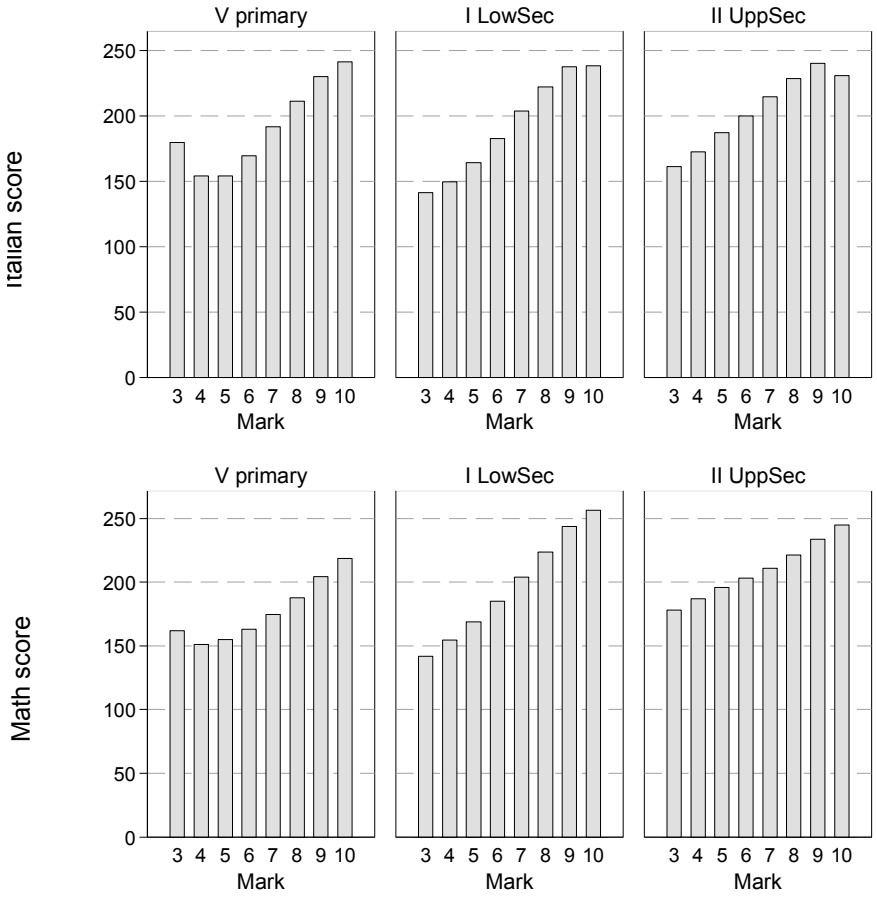


Figure 2. Average test scores in Italian and in Mathematics according to teachers' marks



These data also suggest that although marks in Italian are on average slightly higher than those of mathematics, teachers of this second subject seem to be using a wider range of values for the attribution of judgments compared

to teachers of Italian.

Coherence between marks and test scores

We can move now to the coherence between marks and scores. Looking at how they co-vary, we shall ask: do higher marks reflect higher scores? Figure 2 shows the relationship between the average test scores and marks in Italian (first row) and in mathematics (second row). The graphs indicate that there is on overall positive concordance between marks awarded by teachers and their students' performance in the standardized INVALSI assessments; in other terms, students with higher marks in the mid-term report also display, on average, higher results in the standardized tests. The relationship seems approximately linear especially in the central part of the distribution of marks, while an increase in the tails of the distribution, such as passing from the "9" to "10", is associated with a smaller increase in students' competencies.

Table 1 presents similar information, but with a synthetic indicator, namely the linear correlation between marks and test scores, according to subject and educational level. As showed in previous studies, also in our analyses the correlations are in general not very high, but with a broad variation across educational grades and, partially, subjects.

The largest correlations are found for both subjects in the first year of lower secondary education (between 0.55 and 0.60), while the lowest are in the second year of upper secondary education (less than 0.40). The fifth grade in primary education shows instead inconsistent values across the subjects: the correlation between teachers' marks and students' performance in the INVALSI tests is much higher in Italian than in mathematics. In the second part of the table, we see that the low correlations in the 10^{th} grade are not due to the allocation of students in different tracks, since they remain small even within the different types of school.

We are now interested in assessing whether the correlations just discussed vary according to the macro-region where the school is located. From table 2, we can see that the correlation of teachers' marks and INVALSI score is not strongly heterogeneous across geographical contexts, but some differences are in place.

Table 1. Correlation between teachers' marks and performance in standardized tests across school levels (panel 1) and tracks in upper secondary education (panel 2)

	Italian	Mathematics
Panel 1: School grade		
V Primary	0.549	0.381
I Lower secondary	0.560	0.599
II Upper secondary	0.379	0.332
Panel 2: Tracks in II Upper secondary		
Lyceum	0.330	0.322
Technical	0.357	0.337
Vocational	0.344	0.289

More specifically, we found larger correlation in the Northern regions both in primary and lower secondary education, especially if we look at mathematics.

Table 2. Correlation between teachers' marks and performance in standardized tests across geographical areas, school levels (panel 1) and tracks in upper secondary education (panel 2)

Grades	V Primary		I LowSec		II UppSec	
	North	South	North	South	North	South
Italian	0.571	0.503	0.562	0.535	0.331	0.396
Mathematics	0.411	0.313	0.639	0.504	0.319	0.294
Tracks in Upper secondary education	Lyceum		Technical		Vocational	
	North	South	North	South	North	South
Italian	0.307	0.335	0.320	0.294	0.314	0.277
Mathematics	0.320	0.274	0.330	0.245	0.295	0.215

In the 10th grade, however, the picture is not so clear, since the correlation among mathematics marks and scores is very similar in the two macro-areas, while the correlation for Italian is slightly larger in the Southern regions. The North-south gap in the degree of concordance between teachers' marks and their students' test performance is relatively similar in the lyceums for both subjects, while it is larger among technical and vocational schools, where teachers' marks in mathematics are very lowly correlated with students' test scores in the Southern regions.

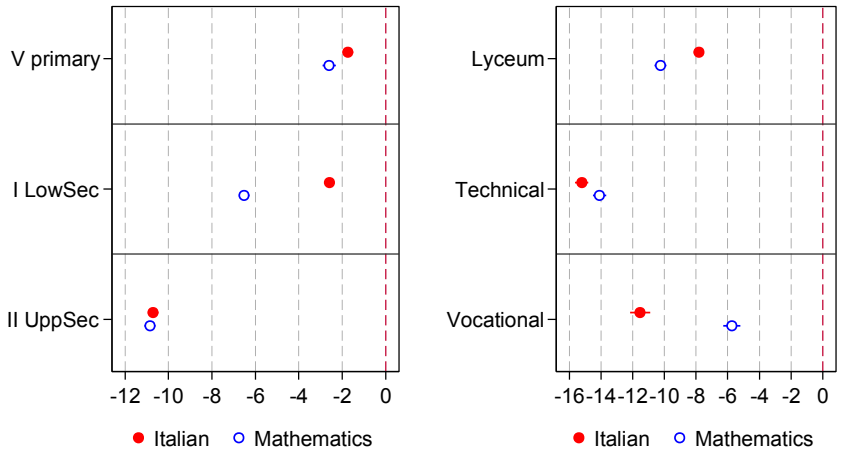
Grading standards

To investigate the generosity/severity and the relevance of students' performance in the attribution of marks to students, we ran separate OLS linear regressions for the three school grades, testing whether there are differences in the INVALSI scores among macro-regions, once that they are (statistically) equalized regarding the marks and the distribution of students' socio-demographic characteristics. Results are reported in graphical form, in order to facilitate the comparison of results across subjects and educational levels. Figure 3 reports the average difference between the South (dots) and the North (omitted reference group, dashed line), by subject (full dots for Italian and hollow circles for mathematics) according to school level (left) and track (right).

We detect that, for each school level and track, the performance of students with the same teacher's mark in Southern regions is lower than in Northern regions. Following the literature on this topic, this can be read as a sign showing the lower severity of teachers in grading students in Southern regions or, as we argued above, as a sign that teachers in Southern regions tend to attribute marks also on the basis of characteristics other than the mere student subject competencies. Nonetheless, there is another possible explanation: Southern teachers could be biased in assessing their students because ranking them within different geographical skills levels. Namely, the observed bias could be due to the fact that grades are attributed to relative performance in classes and schools and not on the basis of a unique national standard. This is likely to happen when teachers tend to "grade on a curve", implicitly adopting a *norm referenced evaluation* (assess the student performance in relation to

those of the other pupils of the same class) instead of a *criterion referenced evaluation* (reward the absolute level of competencies of each student) (Lambert and Lines, 2001).

Figure 3. Linear OLS regression model: average difference between the South and North in the predicted test scores, teachers' marks being equal, according to subject (symbols), school level (left) and track (right)

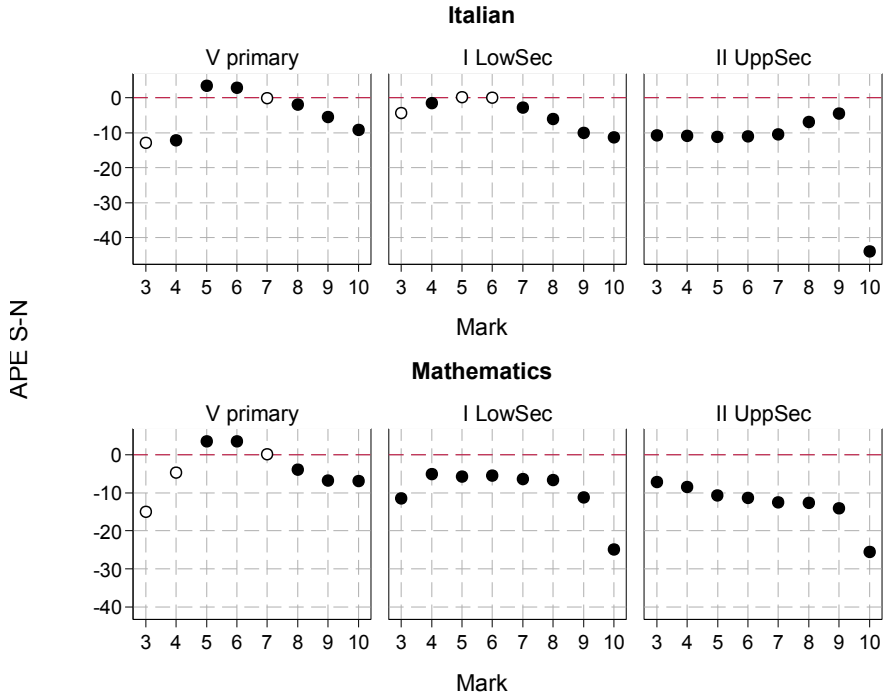


However, it is important to assess whether this gap is homogeneous or systematically varies across educational levels and subjects, since this information is still lacking in the Italian context. First of all, looking at reading-Italian, the difference between the two macro-regions is smaller among students in the fifth and sixth grade, while it is larger in upper secondary education.

The overall pattern for mathematics is similar; the distance between the geographical areas is slightly larger for this subject in primary and upper secondary education, while it is remarkably bigger in lower secondary education. In more concrete terms, considering two 2nd year students in upper

secondary education, with the same socio-demographic profile and with the same mark in mathematics in the mid-term report, we can expect that the one who is studying in a Southern region will have a performance in the corresponding test score that is 11 points lower compared to a similar student in the Centre/North.

Figure 4. Linear OLS regression model: average difference between the South and North in the predicted test scores (teachers' marks being equal), according to marks and subject



Within upper secondary education, there is some heterogeneity across tracks, even if it is of modest magnitude. The largest gap, at a disadvantage of

the South, is found among technical schools in both subjects. Looking at the lyceums, teachers in the South seem to adopt particularly less strict standards in mathematics, while in vocational schools the lower standards are mainly visible in Italian.

Figure 4 shows the results of our second series of regression models, where we interacted the dummy for the macro-geographical area with marks, in order to assess whether the discrepancy in the level of teachers' grading varies along the distribution of marks. We can see that the values corresponding to the gap between Southern (dots) and Northern regions (dashed line) are mainly under the axis and that the distance tends to be larger for the higher marks. Therefore, this figure suggests that the teachers employed in schools located in the South and Islands are particularly prone to be generous when assessing high performers or to incentive them with marks less linked to their actual subject knowledge. This result can also be thought to be responsible of the smaller correlations between teachers' marks and students' test scores in the South, described in the previous section.

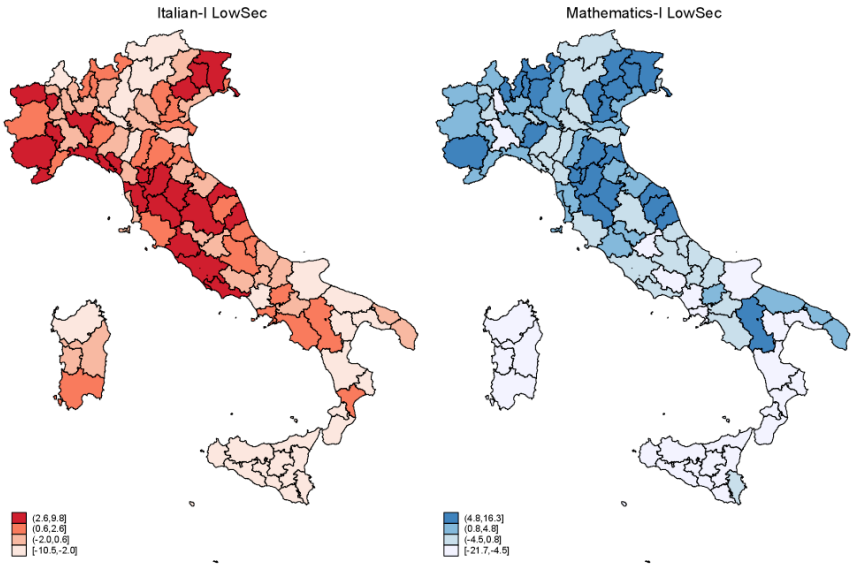
The last part of the analysis is devoted to assess the geographical variations in grading standards using a more fine-grained classification of geographical contexts. More specifically, the data allows us to examine for the first time in our country variations in grading standards at the province level. They are measured as average differences between the predicted test score for each province and the national average, controlling for teachers' marks and student socio-demographic composition. Given the high number of provinces, we report the results using coroplete maps (figure 5) reporting the relative degree of severity in teachers' grading compared to the national standards, expressed in quartiles.[8] The darker colours indicate provinces with teachers who adopt on average stricter standards than national ones, while lighter colours indicate more generous standards. The first map with red colours refers to Italian, the second with blue scale refers to mathematics. Even if we conducted analyses for each educational level, we present in this work only those for the first year

[8] The maps were drawn using the user-written command -spmap- (Pisati, 2008) within the statistical software Stata 13.1.

of lower secondary education[9], as exemplificative of this type of exercise.

As expected, we can see that overall Southern provinces display lighter colours, suggesting that their teachers adopt less severe standards in evaluating students or, at least, attribute less relevance to students' performance in the formulation of their marks.

Figure 5. Coroplete map of grading standard in I lower secondary education, in Italian (left) and mathematics (right)



Note: darker colours indicate more severe standards, while lighter colours indicate more generous standards

Nonetheless, we see also that there is a relatively high heterogeneity among provinces within macro-regions and that the picture is far away from a sharp divide between Southern/Northern standards. Furthermore, the maps also show that there is variability within the same region, both in the North/Centre and in

[9] Due to space constraints we present only the map for lower secondary schools, but the general picture is pretty similar for primary schools and upper secondary schools.

the South/Isles. For instance, we can see that in Sardinia there are provinces with higher standards in Italian/reading than the national average, such as Cagliari (II quartile), but also others with lower grading standards, like Sassari/Olbia-Tempio. The same pattern can be found in Lombardia, where there are both provinces belonging to the quartile with higher standards (e.g. Pavia) and to the lowest one (e.g. Brescia). Finally, it is also interesting to note that there is a relatively high degree of correlation between the grading standards in Italian and mathematics at the province level in the sixth grade, amounting to 0.70.

Discussion and conclusions

In this article we compared students' marks and test scores, having the opportunity to reflect about marks as a signal among different regions of Italy. As we stated, comparing marks and scores is widely used but problematic. Nonetheless, from this comparison, it is possible inferring how grading standards and marks' meaning differ among Northern and Southern regions. Moreover, thanks to the INVALSI national assessment, our analyses have the opportunity to develop a systematic comparison across school levels and subjects and to disaggregate the analyses among upper secondary tracks and provinces. We worked on differences in grading coherence, looking at the correlation between marks and scores, and in grading generosity/severity, comparing the differences in predicted scores for students with the same mark.

We observed that at the national level there is a positive correlation between marks and scores and this is stronger in primary (but not for mathematics) and lower secondary education, while it is weaker for all tracks in upper secondary education. This indicates that teachers' evaluation move in the same direction of the results obtained by their students in the standardized assessments.

Coherently with previous studies, our analyses showed that Southern regions are characterized by what seems higher generosity in grading students. Indeed, we detected that Southern students display lower performance in the INVALSI assessment compared to Northern students with the same marks and socio-demographic profile. We interpret this gap as generosity because, as we

noticed, the correlation between marks and scores is not so different among Northern and Southern regions. This result suggests that the higher marks received by Southern students are not due to lower weight attributed by their teachers to subject competence. Moreover, we noticed that the distance between marks and performance is more intense on high performing students. In other words, the generosity bias is larger especially for higher marks (9 and 10), hence stronger on the right tail of the distribution.

Comparing grading standards in mathematics and reading-Italian on the same population of students gave us the opportunity to detect that the generosity bias is stronger in mathematics than in reading[10], especially in lower secondary schools and lyceums.

Our findings consolidated and enriched the general finding of more generous grading standards in Southern regions, previously highlighted in Italian research (IReR, 2006 & 2008; Gay & Triventi, 2011) . At the same time, our research confirms how much grading standards are heterogeneous among schools and contexts in a novel way: indeed, we observed that there are striking differences in grading standards among Italian provinces even within macro-regions.

Therefore, the North/South divide is a relevant interpretative key, but it is also partial and risks being misleading. We mean that the North/South dichotomy hides a relevant fact, the high heterogeneity existing among provinces and schools in the generosity/severity bias. The implications of the heterogeneity in grading standard among Northern and Southern regions could be quite severe. Indeed, as we argued, marks are a crucial signal that students and their parents receive from teachers. Over-rating Southern students risks distorting the information used by students and families in taking educational choices, especially at the end of lower secondary school, a critical point in the Italian educational system.

Finally, it should also be considered that marks obtained by students in upper secondary schools could have relevant implications: they could be used by employers in the labour market and by universities in the selection of

[10] The only exception to this result is upper secondary vocational track.

applicants. Due to these reasons, the fact that marks are poorly correlated to students' skills in upper secondary schools is quite worrying.

To conclude, the fact that grading standard are far away from being uniform among country regions is a result that should be further investigated. In our opinion an important additional step will be moved once that data collected on students is matched with information coming from their teachers. Our suggestion is to explore the mechanisms leading to generosity/severity biases and to low correlation between marks and scores, in order to understand whether and how educational policies could tackle the heterogeneity in grading standards.

Acknowledgements: The authors would like to thank INVALSI for allowing us use the data for this publication. We would like to thank the participants of the VII Espanet Conference (Turin, 18-20 September 2014) and the Final Seminar of the "Concorso idee per la ricerca - Invalsi" (Rome, 9-10 December 2014), as well as the anonymous reviewers of the journal for useful comments and suggestions. All errors remain ours.

References

Argentin, G., & Tamanini, G. (2001). Il nuovo esame e le opinion degli studenti. In C. Tamanini and C. Tugnoli (Eds.), *Gli esami di Stato in Trentino: Ricerca e Laboratori* (pp. 41-72), IPRASE-Studi e Ricerche.

Betts, J. R., & Grogger, J. (2003). The impact of grading standards on student achievement, educational attainment, and entry-level earnings. *Economics of Education Review*, *22*, 343–352.

Bolletta, R. (2001). *Studio sperimentale sull'assegnazione dei punteggi nelle prove scritte.* Istituto Nazionale per la Valutazione del Sistema dell'Istruzione, Osservatorio Nazionale sugli Esami di Stato.

Bonesrønning, H. (1999). The variation in teachers' grading practices: causes and consequences. *Economics of Education Review*, *18*, 89–105.

Bonesrønning, H. (2004). Do the teachers' grading practices affect student achievement? *Education Economics*, *12*(2), 151-167.

Bonesrønning, H. (2008). The effect of grading practices on gender differences in academic performance. *Bulletin of Economic Research, 60*(3), 245-264.

Bowles, S., & Gintis, H. (1976). *Schooling in Capitalist America*: Educational Reform and the Contradictions of Economic Life. New York: Basic Books.

Bratti, M., Checchi, D. & Filippin, A. (2007). *Da dove vengono le competenze degli studenti*. Il Mulino: Bologna.

Checchi, D., Bratti, M., & Filippin, A. (2011). *Progetto n. 1 Valore di segnalazione del voto di diploma e grading standard nelle scuole secondarie superiori.* Working Paper of Centro interdipartimentale "Work, Training and Welfare Department", University of Milano.

Costrell, R. M. (1994). A simple model of educational standards. *American Economic Review*, *84*, 956-971.

Dardanoni, V., Modica, S. & Pennisi, A. (2009). Grading across schools. *The B.E. Journal of Economic Analysis & Policy*, *9*(1), Article 16.

Gasperoni, G. (1998). *Il rendimento scolastico*. Bologna: Il Mulino.

Gay, G., & Triventi, M. (2011). Voti in italiano e competenze in lettura: come variano gli standard valutativi in Italia?, In U.S.R. (Ed.), *Le competenze degli studenti lombardi. Il rapporto OCSE-PISA 2009 in Lombardia: risultati ed approfondimenti tematici* (pp. 143-165), Brescia: Vannini.

Guskey, T. R. (2000). Grading policies that work against standards and how to fix them. *NASSP Bulletin*, *84,* pp. 20-29.

Johnes, G. (2004). Standards and grade inflation. In G. Johnes & J. Johnes (Eds.), *International Handbook On The Economics Of Education* (pp. 462-483). Cheltenham, UK: Edward Elgar.

Iacus, S. M., & Porro, G. (2011). Teachers' evaluations and students' achievement: a 'deviation from the reference' analysis. *Education economics*, *19*(2), 139-159.

INVALSI (2010). *Le competenze in lettura, matematica e scienze degli studenti quindicenni Italiani. Rapporto nazionale PISA 2009.* Roma: Istituto Nazionale per la Valutazione del Sistema Educativo di Istruzione e Formazione.

INVALSI (2012). *Rapporto nazionale sulla rilevazione degli apprendimenti 2011-12*. Roma: Istituto Nazionale per la Valutazione del Sistema Educativo di Istruzione e Formazione. Retrieved from http://www.invalsi.it/snv2012/documenti/Rapporti/Rapporto_rilevazione_apprendimenti_2012.pdf

IReR (2006). *Valutazione degli apprendimenti disciplinari nella scuola secondaria di primo grado.* Rapporto finale. Milano: Istituto di Ricerche della Regione Lombardia.

IReR (2008). *La misura del 'merito scolastico' ed i suoi effetti sulle politiche di sostegno del sistema educativo di istruzione e formazione*. Rapporto finale, Milano: Istituto di Ricerche della Regione Lombardia.

Koretz, D. (2008). *Measuring Up*: What Educational Testing Really Tells. Cambridge, MA: Harvard University Press.

Lambert, D., & Lines, D. (2001). *Understanding assessment. Purposes, perceptions, practice.* Taylor&Francis e-Library.

Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment. *Journal of Public Economics*, *92* (10–11), 2083–2105.

Lillard, D. R., & DeCicca, P. P. (2001). Higher standards, more dropouts? Evidence within and across time. *Economics of Education Review*, *20*, 459–473.

OECD (2012). *Grade expectations: How marks and education policies shape students' ambitions*. Paris: OECD Publishing.

OECD (2013). *Grade expectations.* PISA in Focus no. 26. Paris: OECD Publishing.

Pisati, M. (2008). *SPMAP: Stata module to visualize spatial data*. Retrieved from: http://EconPapers.repec.org/RePEc:boc:bocode:s456812.

Polloway, E. A., Epstein, M. H., Bursuck, W. D., Roderique, T. W., McConeghy, J. L., & Jayanthi, M. (1994). Classroom grading. A national survey of policies. *Remedial and Special Education, 15*(3), 162-170.

Sestito, P., & Tonello, M. (2011). *I differenziali nella qualità degli iscritti alle Università Italiane: il caso delle Facoltà di Medicina e Chirurgia*. Questioni di Economia e Finanza (Occasional Papers), Banca d'Italia.