UNIVERSITÀ DEGLI STUDI DI TRENTO
Dipartimento di Sociologia
e Ricerca Sociale

School of
Social Sciences

WHEN THE TEACHER MARKS THE DIFFERENCE

GRADING PRACTICES, SOCIAL INEQUALITIES AND STUDENT EDUCATIONAL OUTCOMES

Ph.D. Candidate:
Ilaria Lievore

Supervisor:
Prof. Dr. Moris Triventi, University of Trento

Doctoral Committee:
Dr. Sara Geven, University of Amsterdam
Prof. Dr. Ruud Luijkx, University of Tilburg
Prof. Dr. David Reimer, University of Aarhus

Academic Year 2021/2022

Dissertation submitted to the Department of Sociology and Social Research at the University of Trento, within the Ph.D. program in Sociology and Social Research, Doctoral School of Social Science.

TABLE OF CONTENTS

ACKNOWLEDGMENTS

First, I want to thank the doctoral committee: Dr. Sara Geven, Prof. David Reimer and Prof. Ruud Luijkx, for taking the time to read my dissertation and for the useful comments and insights. I really appreciate your effort and commitment. I want also to thank my Ph.D. supervisor Moris Triventi for accepting to mentor me (twice!) and for the countless discussions, revisions, brainstorming. Thank you for the patience and for the Stata tricks, you taught me a lot.

Thanks to my colleagues from the 34th cycle, and to all the other colleagues of previous and following cohorts from Trento, and a special thanks to the Edu Team: you are truly the best support I could have hoped for.

I want to thank all my closest friends who supported me during these last 4 years, in particular Chiara, Elena, Giulia, Irene, Vale, Vero, Vesna: thank you for being always by my side.

I want to thank my family. My sister Elena, thanks for believing in me even when I couldn't: you are my everyday inspiration. Mamma e papà, grazie per non aver mai messo in discussione le mie scelte e di esservi sempre fidati di me, dandomi sempre un posto sicuro dove poter tornare: è il regalo più grande che possa ricevere sempre.

The last thank goes to Mauro. Thank you for seeing me, supporting me, and caring for me. You are my home, and I will always find a way to come back to you. Thank you for making this Ph.D. worth it.

INTRODUCTION

The expansion and the reformation of the educational system in Western Countries in the last century led to a significant increase of the average level of educational attainment as well as an increase in the differentiation in the educational systems. Nowadays, the need for equality in educational opportunities is universally recognized, especially considering a comprehensive diffusion of knowledge and basic literacy and numeracy skills as one of the most important goals in present societies. Nevertheless, empirical research has consistently shown significant patterns of inequalities in educational attainments, correlated to the belonging of different social groups (Barone & Ruggera 2017). Indeed, in educational research, even if slowly declining, patterns of educational inequalities linked to students' ascriptive characteristics, such as gender, socioeconomic status, and ethnic background, have been explored and acknowledged all around European countries (Breen et al. 2009), and students' ascriptive characteristics have been found to exert significantly strong and durable effects on children's educational attainment in terms of academic competences and educational choices. Still, there are several open questions concerning the underlying mechanisms of the reproduction of educational inequalities, and regarding the effect on a number of students' outcomes concerning their educational attainment.

The most important theoretical contribution in the broader framework aiming to explain the complexity of the mechanisms fostering educational inequalities is historically attributed to Boudon (1974). The author managed to define inequalities in educational opportunities as the consequence of two specific effects: social-background differences in academic performance are due to primary effects, while social-background differences in the choices that students and parents can make are due to secondary effects. Primary effects (or performance effects) are due to differences in the academic performance across social groups, given by a complex interaction between educational institutions and cultural, economic and social resources of students and their parents. Factors that generate primary effects are summarised as: a) genetic; b) home environment and economic, cultural and social resources; c) health and nutrition; d) sibship size, e) cultural biases exhibited by institutions and f) psychological mechanisms (Jackson 2013). Secondary effects (or choice effects), instead, explain how costs, benefits and expected probabilities of success associated with academic outcomes vary according to socio-economic background. Even when children have similar competences, social background affects educational choices, i.e., students from low social background with high academic performance are however more likely to make less ambitious educational choices than their more privileged counterpart. (Alexander, Entwisle & Olson 2001). Secondary effects are due to differences in economic resources, educational and career aspirations, norms and values and other macro-socio-economic factors, such as available resources, incentives and institutions.

In this broader picture, many authors questioned about the role of schools, and of teachers, as external influences in the creation and in the reproduction of inequalities in educational opportunities. Teachers are considered to be very influential for students' ability and educational achievement, and regards the process of allocating students in different school tracks (Reimer, 2019). However, little is known about whether and how teachers can influence the accumulation or the compensation of educational inequalities. Teachers play a complex and mostly unconscious role in the reproduction of social inequalities which cannot be ascribed only to primary or secondary effects. For this reason, Scheinder (2014) first, and then Esser (2016) defined as *tertiary* effects the additional effects of social class on the tracking process through the school context and through the influence that teachers have on students. Teachers' attitudes vary according to children's socio-economic background, and this has a direct consequence both on students' educational achievement and school choices (Esser & Relikowski 2015). Tertiary effects, complementary to primary and secondary effects, are supposed to capture inequalities in educational attainment – measured mainly through stratified educational choices – due to institutional sorting processes in which teachers have an active role when showing different expectations according to students' different social backgrounds (Thys 2018). The relatively recent introduction and definition of tertiary effects is analytically important since it broadens the range of actors considered when studying mechanisms of reproduction of educational inequalities, focusing also on school and especially on teachers rather than only on families, considered the pivotal actors for both primary and secondary effects (Argentin & Pavolini 2020).

Sociological literature addressing tertiary effects with a quantitative approach has been grown only in the recent years, and most of the present studies are either experimental or observational. Experimental research addressing teachers is mostly present in the field of social psychology (Geven et al. 2018), while observational data are used by sociological researchers addressing whether and how teachers treat students with different social backgrounds, gender or ethnicity differently (Reimer 2019). The ways in which these differences in treatment manifest can be related to: teacher expectations, as prediction of students' future academic performance; teacher grading and teacher judgments, as teacher assessment of a current academic performance (Rubie-Davis et al.

2015); teacher recommendations and advices concerning students' future educational choices (Boone & Van Houtte 2013). However, the number of recent sociological studies that address the role of teachers with this perspective is inadequate in order to give a comprehensive view of the phenomenon (Reimer 2019).

The present dissertation argues that, along with quantitative analyses on primary and secondary effects, studying teacher effects and tertiary effects can help improving knowledge about the complexity of mechanisms interplaying in the reproduction of inequalities in education. Thereby, the aim of this dissertation is to gain a more accurate understanding of the mechanisms that come into play when considering the role of teachers, considering teacher effects from different perspectives, and considering both primary and tertiary effects. Regarding tertiary effects, I propose a broader definition of as compared to the definition provided by Esser (2016). Indeed, when looking at the role of teachers, I do not limit the focus on differentiations in expectations, attitudes, and evaluation according to students' social class and socioeconomic background. Tertiary effects are conceived also regarding variations in teachers' expectations according to students' ascriptive characteristics more generally – social background as well as gender and migratory background. Moreover, tertiary effects are not considered as affecting only students' educational decisions due to sorting processes, but teachers' evaluations are considered as of the main channels affecting students' educational attainment in terms of both performance and choices, when they are shaped by teachers' expectations and teacher biases. Tertiary effects have indeed direct consequences on the allocation of students in different educational tracks (Esser 2016; Thys 2018) but they might also have indirect consequences through teachers' evaluation of students' competences and abilities. Tertiary effects are thought as capturing differences in teachers' evaluation associated with different expectations and unconscious stereotypes associated to students' belonging to specific social groups, which may lead to biased grading according to students' ascriptive characteristics. However, it is important to separate the effect of teachers' characteristics and teacher quality on students' educational attainment, which is still conceived as primary effects, even if belongs to the bigger concept of "teacher effects". Therefore, while acknowledging the fundamental role of schools as organizations in shaping tertiary effects, the focus is on the role of teachers, teacher effects connected to primary and tertiary effects, and on the micro-mechanisms explaining how teachers may interact with the educational system reproducing at the micro-level social inequalities in educational outcomes.

The analysed context is the Italian educational system. Italy is an interesting country for studying tertiary effects and teacher grading practices for several reasons. Its education system is almost entirely public-funded, especially concerning compulsory schooling – primary and secondary education. In the last 20 years, the Italian educational system has seen several reforms promoting decentralization and school autonomy. However, they produced contradictory effects (Grimaldi & Serpieri 2012). Indeed, despite the reforms, the education system is regulated by formal processes with the aim of providing a relatively uniform educational environment (Argentin & Pavolini 2020) through a supposedly equal allocation of resources among schools. Schools are not entirely autonomous in their administration choices, therefore assuring some geographical stability and harmonization concerning the basic organization of schools, institutional features and curricula.

However, even if formally the educational environment is homogeneous, the Italian education system is characterized by vast territorial divides in school-resources and socio-cultural environment (Montanaro 2008). Regional differentiations as well as local differentiations shape the Italian system, in which segregation is likely to happen both between schools (fight for "better schools") and within school (sorting of students among different classes) (Argentin & Pavolini 2020). In this picture, the teaching profession in Italy is granted by the constitutional right of "autonomy" in duty delivery. Consequently, teachers in compulsory schooling have a considerable degree of autonomy and independence from other teachers and even the school administration (Bracci 2009) in deciding their exam structure, the frequency of their evaluation and especially their grading practices. The degree of standardization and centralization of exams and assessments is very low (Pensiero, Giancola & Barone 2019), and the territorial heterogeneity in grading standards is pretty high (Argentin & Triventi 2015). Above all, the Italian system show high levels of social inequalities with regard both academic performance and tracking, which are stratified in each grade of compulsory schooling according to students' gender (Bozzano 2012; Borgna & Struffolino 2017), ethnic background (Azzolini & Barone 2012; Azzolini, Mantovani & Santagati 2018; Pensiero, Giancola & Barone 2019) and social origin (Contini & Triventi 2016; Ballarino, Panichella & Triventi 2014; Barone & Ruggera, 2018). For these reasons, Italy con be considered a perfectly suitable scenario in which heterogeneity in grading bias, and grading favouritism, is likely to occur.

In Italy, every year the national bureau for school testing (INVALSI) carries out assessment of students' academic competences along compulsory schooling in grade 2, 5, 8, 10, and recently also grade 13 has been included in the assessment. These assessments include the whole population of Italian students belonging to the mentioned grades, are standardized at the national level, and the results are nor communicated to students, teachers or families – so they are not part of the final evaluation of students. For some grades in specific academic years, also students' and teachers' questionnaire are administered, with the possibility of linking information of school and teachers to students. This important feature of the INVALSI-SNV dataset makes it well-suited for investigating the role of teachers in the reproduction of educational inequalities in the Italian system, and the use of Italian student-teacher matched data in this dissertation represents a novelty in educational research. Moreover, INVALSI-SNV data allows me to: focus on different educational levels, from primary school to upper secondary school; have information on the whole student population; have information on teachers' grades together as well as on students' competences measured through a standardized test; rely on rich information on both students' sociodemographic characteristics as well as on teachers' characteristics; and finally, follow students through their academic career thanks to a unique code with whom students are identified once they enter their educational journey.

To sum up, this dissertation aims at providing a general theoretical framework through which teacher effects can be analyzed through a sociological perspective. This framework accounts for 1) tertiary effects, defined as the impact teachers have on students' educational decisions, also indirectly via students' evaluations, according to their variation in expectations and biases related to students' sociodemographic characteristics; 2) primary effects related to teachers, defined as the impact teachers have on students' performance according to their grading preferences, quality, characteristics. The empirical chapters provide three different examples allowing the study of teacher effects using different perspectives. In the following paragraphs, each chapter is introduced. I provide more specific research questions and a summary of the conclusions derived from each empirical work developed. In chapter 1, I propose a general theoretical framework and a broader definition of tertiary effects aimed at introducing the micro-mechanisms according to which teachers interact in the educational system and may shape inequalities in educational outcomes with respect to students' gender, social status and migratory background. In this framework, it is underlined how the role of teachers and the interaction between teachers and, respectively, students, parents, and schools as institutions are multifaceted and complex. Many aspects come into play altogether, and as quantitative research the challenge is disentangling the effect chain as regard to teacher effects. Therefore, chapter 1 aims at proposing a comprehensive description of the set of mechanisms that intervene in the relationship between teachers and students' educational achievements on one side, and students' school choices on the other side. Looking at the relationship between students' characteristics and their academic performance, teacher quality, teacher characteristics, as well as teacher expectations and grading practices may come into play. As a result, grading bias according to students' ascriptive characteristics may occur, meaning that students with the same ability and competences measured through standardized tests may have different academic performances measured through grades. This in turn may influence how teachers allocate students into different tracks through recommendations, which can also be biased according to students' gender, ethnicity and socioeconomic status. Therefore, the multifaceted role of teachers, and how they impact the accumulation, or the compensation of inequalities is questioned.

Chapter 2, 3 and 4 provide empirical examples framed in the sociological literature about tertiary effects, relying on quantitative methodological approaches for the study of teacher bias. The common methodological framework of the empirical chapters relies on the study of grading practices, as a comparison between an "objective" and unbiased measure of students’ competences, such as standardized test scores, and teacher grades, which are thought to be more subjective and possibly biased measure of students’ academic ability. This is accomplished using an economist perspective in Chapter 2, where through the comparison between the two different measures it is possible to obtain some information about teachers grading standards. In Chapter 3 and 4 instead I adopt a sociological perspective, using the grade equation model, in which teacher grades are expressed as a function as a variable identifying the group of interest – such as gender, ethnicity or social status – plus the “objective” measure of student academic ability. The comparison between the two measures, teacher grade from one side, and student score in a standardized test from the other side, may provide the extent to which teachers are likely to reward or penalize students from different social groups.

The goal of the second chapter is to capture a causal effect of the adoption of specific grading practices on students’ academic competences and educational choices later. After creating a measure of teacher grading standards, accounting for how much the teacher is strict rather than generous when assigning grades to their students, I rely on an instrumental variable approach to determine whether higher grading standards measured in primary schools have an effect on students’ educational outcomes measured later on in time. A sample of 9,370 students matched with their teachers in 5th grade is followed up to 8th and 10th grade. Results demonstrate that students with a stricter teacher in 5th grade, therefore with higher grading standards, have higher performance in both Language and Mathematics in 8th and 10th grade, and are more likely to be enrolled in the academic track in 10th grade. Interestingly, the positive effect of higher grading standards is pretty stable among students with different gender and migratory background, and belonging to different social classes, with the exception of immigrant students and their performance in Language measured 5 years later.

The aim of the third chapter is twofold. On the one hand, it provides empirical evidence of the gender grading gap in Italian upper secondary schools. On the other hand, it analyses the role of contextual factors in affecting the extent of the gender grading mismatch. Specifically, according to previous literature, it examines whether teachers grade female students more generously compared to male students who have the same subject-specific competence, as measured via standardized test scores, and examines whether this putative gender grading premium varies according to key teacher characteristics, features of the classroom, and school type. Results from multilevel analyses on a sample of 38,975 Italian students in 10th grade matched with their teachers show that the latter are more likely to grade female students more generously in both Language and Mathematics. This premium in grade is overall stable even when accounting for teachers' characteristics, classroom composition and type of upper secondary school. It is not possible to derive that teachers discriminate a specific gender group over the other, since it was not possible to account for other specific educational signals that may be determined for teacher judgments, such as behaviour in the classroom, participation, effort, engagement and so on. Therefore, what drives the gender grading mismatch might be either actual behaviours adopted by female students or teacher expectations related to gender, or a combination between the two.

In chapter 4 the aim is accounting for students' socioemotional skills in influencing teacher judgments, net of students' ascriptive characteristics such as gender, migratory background and social status, and students' academic performance. This is accomplished by relying on a novel dataset that merges information on 15-years old students from the INVALSI dataset with the PISA 2018 dataset. Analyses performed on a sample of 6,504 students show how they can be partitioned in three profiles according to the distribution of their socioemotional traits and attitude toward school. Belonging to the profile with the lower performing students in terms of non-cognitive skills has a significant and stable detrimental effect in terms of grade, over and above ascriptive characteristics and performance. Interestingly, gender and ethnic background seem to be important grade determinants also when controlling for the rich set of socioemotional skills provided by PISA. The relationship between ascriptive characteristics and grade in Language is not moderated by student's socioemotional skills, while for grade in Mathematics, it seems that socioemotional skills associated with specific social groups might have a stronger correlation, questioning the common belief according to which grading in mathematics are less prone to bias.

Overall, this dissertation aims at investigating some micro-mechanisms involving teachers may contribute to the reproduction of social inequalities in the education systems related to gender, ethnic background and social class. Results suggest that the role of teachers is strong, and it affects both students' academic performance and educational choices, even in an open educational system such as the Italian one, where teacher grades and recommendations are formally not binding in accessing specific high-school tracks or University courses. More reflection about the overall message of this dissertation is found in the conclusion section.

CHAPTER 1

WHEN THE TEACHER MARKS THE DIFFERENCE: A THEORETICAL BACKGROUND

Abstract

This theoretical chapter explores the complex set of mechanisms that intervene in the relationship between teachers and students' educational achievements from one side, and students' school choices from the other side. First, a broader definition of tertiary effects is proposed, as the result of two paths: teacher impact on student academic performance, measured through grades, and teacher influence on student educational decisions. Several micro-mechanisms are identified. Teacher quality and characteristics, and teacher expectations and stereotypes may result in grading bias according to students' ascriptive characteristics such as gender, ethnicity and socio-economic background, therefore students with the same ability and competences measured through standardized tests may have different academic performances measured through grades. The same micro-mechanisms involved in the relationship between teachers and students' grades may impact also teacher recommendations which are also biased according to students' characteristics. Therefore, the multifaceted role of teachers, and how they impact the accumulation or the compensation of inequalities is questioned.

Keywords: educational inequality; teacher expectations; grading bias; recommendations; stereotypes

Introduction

What does it mean to be a teacher? Teaching and instruction come often with obstacles and contradictions. Especially in Western countries, where knowledge is diffused and secularized, the educational task is particularly delicate, and sometimes vague and uncertain. Teaching is a complex responsibility in which equilibrium between several mission is needed. Teachers have to transmit basic skills (e.g., literacy and numeracy) while objectively selecting gifted students to which they must transmit more specialist skills. At the same time, they have to contribute to the growth of the personality structure of students, helping them to discover their interests, abilities and who they want to become. In other word, teachers do more than select and socialize pupils, but it is difficult to conceptualize, and it is hardly considered when explaining the role of teachers.

According to Wilson (1962), the role of teachers is particularly hard to define since it involves the precise task of social selection together with the provision of personal services. This means that teacher, as a *social selector*, must allocate students within the educational system in an objective and rational way, looking at potential intellect and knowledge. At the same time, however, the role of teacher as a *socializer* implies "motivating, inspiring and encouraging [children], transmitting values to them, awakening in them a respect for facts and a sense of critical appreciation" (22:1962). However, since there are no clear instructions on how to perform these tasks, there is much more autonomy and discretion, together with lots of uncertainties. Moreover, teaching it is not only providing a "one-way" service from teacher to students and families, but it is a two-way, lasting, and slow process that implies enduring and mutable relationships, especially with the students. For this reason, part of the definition of "being a teacher" implies the dimension of affectivity. A strictly professional, apathetic behaviour is simply not possible, since children need an affectional and emotional context in which they can learn and create a sense of identity. Accordingly, there are six categories in which all the conflicts related to the role of teacher are summarised:

> "(i) Those inherent in the role because of its diverse obligations; (ii) those which derive from the diverse expectations of those whose activities impinge on the role (…); (iii) those arising from circumstances in which the role is marginal; (iv) those arising from circumstances in which the role is inadequately supported by the institutional framework in which it is performed; (v) those arising from conflict between commitments to the role and commitments to the career-line; (vi) those arising from divergent value-commitments of the role and of the wider society" (Wilson 1962:27).

Despite all the responsibilities implied in the job of teaching, it is not socially acknowledged as particularly prestigious, and a large share of the perceived failure of educational systems is attributed to teachers (Ben-Peretz 2001). Teachers have constantly to deal with conflicting external influences they must consider (such as economic trends, the impact of globalization, the professionalization of teaching, school goals and guidelines) but also internal needs, tensions and experiences (Ben-Peretz 2001). Teachers constantly deal with the fact that education is a political enterprise, socially constructed and influenced by economies and cultures of a particular society (Cochran-Smith 2000). Teachers' role and their job are also affected by political and social debates. For example, if the goal of the public-school system is to guarantee an equal common set of knowledge and skills for all students, there is a contradiction with the sociological view of the main goal of schools – and consequently of teachers – that is the meritocratic differentiation of skills based on students' intellect and ability (Raudenbush & Eschmann 2015). Teachers, as one of the most important players in the educational system, are then supposed to act both as a compensatory force, equalizing the input, and as a differentiation force, because the output is not equalizing by definition. Indeed, students have different academic performances when they leave school, with huge variation in proficiency, that are largely driven by the social origins of their parents.

According to Raudenbush and Eschmann (2015), if we imagine a counterfactual situation in which students do not go to school, then school (and therefore teaching) is a powerful equalizing force, especially in the early stages in which differences in skills according to students' ascriptive characteristics as social origin are relatively small. But students belonging to specific social groups, for example those with a privileged socioeconomic background, benefit more from going to school later than their counterparts. This mean that there could be some mechanisms during the educational path that reinforce the productivity and the capacity of learning of the already more skilled students, e.g., those from higher socio-economic background. Teachers' role is collocated in a peculiar position in this intricate and contrasting set of educational aims. They can act at the same time as sources of reproduction of educational inequalities and as sources of compensation of educational inequalities during the entire educational path. All the mechanisms that are related to how teachers can shape student educational future are related to what have been called "tertiary effects". However, mechanisms related to tertiary effects can have both a compensatory/equalizing power, or a differentiating power, according to the specific situation in which the teacher is involved.

An Inclusive Definition of Tertiary Effects

According to Boudon (1974), in order to analyse how students' background affects their educational outcomes, it is necessary to draw a distinction between students' academic abilities and their educational choices and decisions. The analytical framework proposed by Boudon (1974) conceives inequalities in educational opportunities (IEO) as the consequences of two distinct path. The first one regards social differences in students' academic performance, and they are called primary effects of social origin. The second one regards social differences in students' choices and intentions, holding their academic performance constant, and they are called secondary effects of social origin (Boudon 1974; Jackson 2013). Primary effects are the consequence of a complex interaction between institutions and families, carrying cultural, economic and social resources of pupils and their families. Specifically, factors that generate primary effects are summarised as: a) genetic; b) home environment and economic, cultural and social resources; c) health and nutrition; d) sibship size, e) cultural biases exhibited by institutions and f) psychological mechanisms (Jackson 2013). Secondary effects, instead, are connected to pupils' and parents' choices and decisions related to investment in education, both horizontally (decisions among specific tracks) and vertically (decisions about the level of education to be attained). Secondary effects are due to differences in economic resources, educational and career aspirations, norms and values and other macro-socio-economic factors, such as available resources, incentives, and institutions. Social differences are conceived as driven by parental social background, and the concept of cultural capital (Bourdieu 1974) and cultural resources play a substantive role. Therefore, primary effects are conceived as those explaining the relationship between children's socioeconomic background and their level of academic performance through genetic or socio-cultural factors; secondary effects are conceived as expressed via differences in educational choices according to socioeconomic background, even when children have similar previous academic performance (Jackson et al. 2007).

The distinction between primary and secondary effects is quite consolidated in educational literature, but primary and secondary effects refer mainly to mechanisms happening within the family context. For this reason, an attempt to define an additional type of effect has recently been made: tertiary effects (Schneider 2014; Esser 2016; Argentin & Pavolini 2020) refer to an additional path linking students' social background and their educational attainment in relation with both teachers and the school itself as an institution. Dealing with tertiary effects normally implies the effect that children's social origins exert on variations in teacher attitudes. According to Esser's model (2016), educational sorting into different tracks is the result of three factors: student achievement (primary effect), students' and parents' decisions about the educational choice despite student achievement (secondary effect) and teacher recommendations (tertiary effects) (Thys 2018). Consequently, tertiary effects, as defined by Esser (2016), explain the role of schools and teachers in institutional sorting, by underlining the importance that teacher "expectations, efforts, evaluations and recommendations" (Esser 2016:30) may have on students' educational decisions.

Tertiary effects are conceived only for explaining ability tracking models, and only in relation to students' and parents' socioeconomic background. In this dissertation, I expand the definition of tertiary effects in order to account for broader teacher effects on students' educational outcomes, that are defined not only related to students' tracking through teacher recommendations but also to students' academic performance through teacher evaluations. Indeed, if tertiary effects have direct consequences on the allocation of students in different academic tracks (Esser 2016), they may be important to consider for having also indirect consequences through teachers' evaluations of students' performances. Moreover, teacher recommendations and evaluations differentiate not only according to students' social origin, but also according to other students' characteristics. Therefore, the aim of this chapter is to collect and explain in a single framework all the mechanisms that account for how teachers can have an influence on academic performance and educational choices depending on pupils' social backgrounds as well as on other pupils' sociodemographic characteristics such as gender and ethnic or migratory background. This is achieved by considering for the first time a broader definition of tertiary effects, that includes: i) the relationship that occurs between students' ascriptive characteristics and teacher assessment of their ability, since teacher evaluation of student ability may be a factor that influences directly teacher recommendations; and ii) stratification on educational inequalities along students' social origin, gender, and migratory background, since teacher effects may reveal along those dimensions.

There are several mechanisms that can explain the role of tertiary effects, and therefore the role of teachers, in the reproduction of educational inequalities. Broadly, as mentioned above there are mainly two paths through which tertiary effects are (re)produced:

i) *teacher evaluation* of students' ability, measure through teacher grades/judgments/assessments.

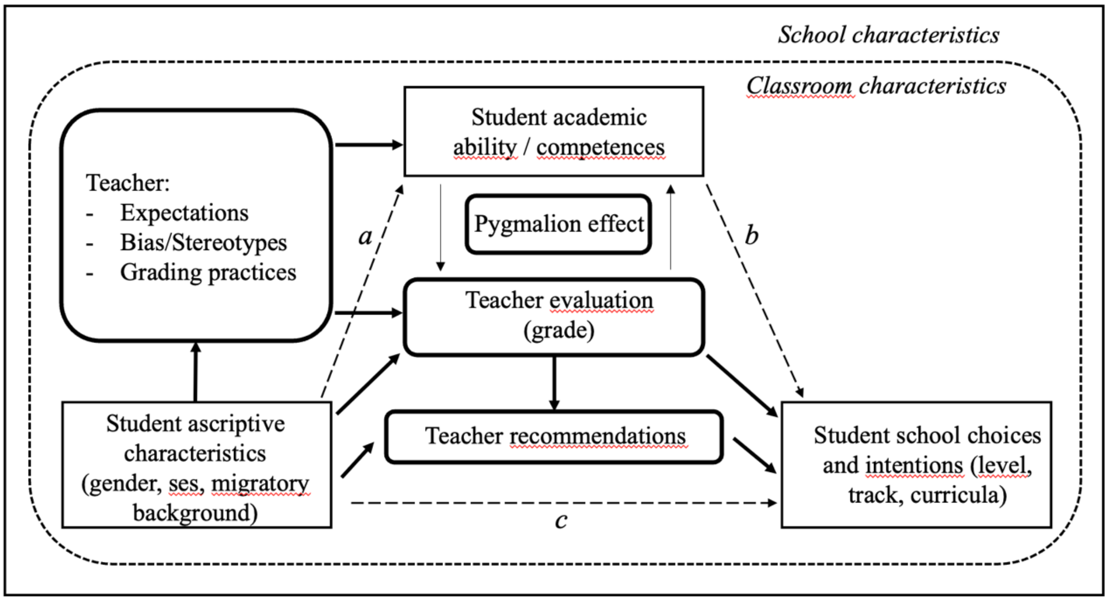
ii) *teacher recommendations*, or influence on student educational decisions.

Figure 1.1 summarizes tertiary effects and the related mechanisms as conceived in this framework, within the primary-and-secondary-effects framework conceptualized by Boudon (1974). Micro-mechanisms explaining how student ascriptive characteristics, student ability, teacher assessment and teacher recommendations are intrinsically interrelated. Teacher quality and teacher characteristics are thought to influence students' academic ability via *primary effects*, while teachers' expectations and bias, and grading practices may lead to social differentiations in teacher assessment in accordance with teacher stereotypes linked to students' characteristics via *tertiary effects*. These dimensions can shape both student ability and student grade, because of the Pygmalion effect (see Rosenthal & Jacobson 1968). Moreover, teacher recommendations, which are affected both by students' ascriptive characteristics and by teachers' assessments of students' abilities, are also linked to teachers' stereotypes (Esser 2016).

An important concept that can be incorporated in the definition of tertiary effects is the definition of "anticipatory decisions" (Jackson et al. 2007). Indeed, educational choices may not be the direct consequence of students' performance. Educational decisions may be elaborated by students and families prior to specific assessment of their performance, in determined key transition points, and these prior decisions may influence the way in which they perform in these tests, mainly through a positive or a negative effect on motivation. Interestingly, also teachers may formulate some "prior decisions" about their students' possibilities in terms of educational choices and evaluate them according to their prior decisions. This is linked with the concept of teachers' expectations.

In this framework, the challenging task is to understand the temporal logic underneath these several dimensions. Indeed, it is fundamental to recall that these processes are intrinsically interrelated, and it is problematic disentangle them, especially because the relationship between students and teachers is built over a relatively long period of time.

Figure 1.1: Conceptual model of tertiary effects and related micro-mechanisms (in bold) within the primary and secondary effects framework


School characteristics
Classroom characteristics
Teacher:
- Expectations
- Bias/Stereotypes
- Grading practices
Student academic ability / competences
Pygmalion effect
Teacher evaluation (grade)
Teacher recommendations
Student ascriptive characteristics (gender, ses, migratory background)
Student school choices and intentions (level, track, curricula)
a
b
c

Note: primary effects = a * b; secondary effects = c. Source: author's elaboration[1].

Teacher Evaluation of Students Ability

In order to properly understand the different teacher effects, it is necessary to specify some concepts as they are used in this chapter. When dealing with student (subject-specific) *competences*, we refer to some "objective" measure of students' ability, such as

[1] I thank prof. Hartmut Esser for the incredibly useful discussion about tertiary effects and about the proposed conceptual schema.

student achievement in standardized tests. Standardized achievement tests are tools that allow comparisons of knowledge and skills of students of the same age or grade in a defined area (Popham 1999), and in educational studies it is also the most accurate proxy of student ability that is available in standard datasets.

Student *grades* refers to teachers' evaluation of student academic ability. In educational studies, there is not always a clear distinction between student performance measured through grades and measured through standardized tests, both from a theoretical and sometimes empirical perspective. Teacher assessment is often considered as the prior indicator of students' competences, but this is not always the case, since teachers' evaluations of students do not take into consideration only their actual competences. Teachers' grades are likely to be imbued with social considerations related to both teachers and students' characteristics, their interaction and the context in which the relationship develops. This finding refers mainly to students' characteristics that they have or show in classroom that may influence teacher perception of their ability. Consequently, different groups of students may perform in a different way in standardized tests score compared to how they may perform in classroom. Therefore, if the two instruments (test score and grades) rank students in different way, it is not clear which of the two serves the purpose of deciding which students should "come first" (Wikström & Wikström 2014).

The way in which teacher assess their students is fundamentally important for several reasons. First of all, they are an indicator for students and parents of the ability of the formers. Indeed, there is a distinction between the real achievement of students – what we called ability and competences – and the perceived achievement – therefore teacher grades. In most of the cases, what really matter to students and families is the perceived achievement. Grades are often determinant in accessing the next level of education, gaining a particular scholarship, or conquering the admission to college (Bonesrønning 2004). That is, grading allows an effective communication of student academic achievement between schools and families, so on one side, students can be efficiently tracked, and on the other side it allows the identification of students who may need additional support (Jalava, Joensen & Pellas 2015). Secondly, the way in which students are evaluated can affect directly their ability and their competences in schools, since grades are an objective measure of teachers' expectations and it can affect the way in which students perceive themselves. This mechanism is usually called Pygmalion phenomenon (Rosenthal & Jacobson 1968). Finally, teacher assessment is used and processed as an important source of information not only by students and families, but also by teachers themselves, who base their educational recommendations on the prior indicator of student ability they have access to, that is, their grades.

Differentiating between grades and competences measured through standardized tests is not straightforward, both empirically and theoretically. These conceptually different dimensions are so related to each other that it is impossible from a causal perspective to understand "what comes first". The same happens when dealing with the relationship that develops between children and their teacher, since, as mentioned early, it is a two-way process in which expectations, ideas and lessons are constantly exchanged during a relatively long period of time. The following sections collect all the mechanisms related to how teachers assess their students depending on their ascriptive characteristics. However, also teacher characteristics, together with student ones, play a fundamental role in shaping the relationship between student and teacher. This is the reason why, given that teacher characteristics is one of the dimensions of teacher quality, the latter is also briefly described. It is important to notice that mechanisms related to tertiary effects are strictly related to one another in causal terms, therefore the conceptual separation of all those dimensions is only for descriptive purposes rather than for describing what really happens at schools.

Teacher Quality and Teacher Characteristics

Teacher quality and teacher characteristics are imbued in the sphere of *primary effects* in shaping educational inequalities. Concepts such as teacher *quality* or teacher *effectiveness* are often difficult to interpret. This is due to the fact that different researchers have used different ways of measuring and detecting the characteristics, habits and practices attributable to a “good teacher”. Nevertheless, there is agreement, and it has been widely demonstrated, that some teachers are more effective than others, and some of them are more able to contribute to their students’ intellectual and personal growth than others. There is almost universal consent that this has consequences in terms of student achievement and ability (Goe 2007). In previous literature there are many different (and sometimes contradictive) ideas when trying to give a definition of teacher quality. However, there are overall three broad elements to which we can refer to for establishing what a “good teacher” is: teacher *qualifications*, teacher *practices* and teacher *characteristics* (Lewis et al. 1999, Goe 2007, Stronge et al. 2011).

Teacher *qualification* refers to all the resources that are considered important in “establishing who should be allowed to teach” (Goe 2007:10), such as educational degree and certification, professional development and seniority. Teacher *practice* refers to what teachers do in their classrooms, the ways in which they decide to operate and the learning methods they adopt. Some of the dimensions of teacher practice are instructional delivery and differentiation, complexity, questioning, and choices of test format. Finally, teacher *characteristics* include attributes, attitudes and beliefs of teachers that are hard to change almost as much as ascriptive characteristics such as gender, ethnic and social origin. Considering teacher characteristics as a component in determining teacher quality means that, consequently, there are some components of teacher quality that are "logically, ethically, or practically beyond the teacher's (or school's) ability to change" (Goe 2007:10). Indeed, not considering such characteristics would imply overlooking a dimension which is as important as the others in determining student academic and educational success.

Concerning ascriptive characteristics of teachers, scholars suggest that an increase in achievement is verified when there is student-teacher match. For example, if we consider ethnic origin, children assigned to teachers of the same race will benefit in terms of ability measured in standardized test score (Egalite et al. 2015; Pitts 2007; Clotfelter, Ladd & Vigdor 2007; Dee 2004). Minority students can benefit from assignment to teachers of their same race and ethnicity because these teachers can serve as model and mentors (Egalite et al. 2015). Considering student-teacher gender match, there are controversial results about the presumption that female teachers have better relationship with girls (who therefore benefits in terms of ability) than boys and vice versa (Spilt, Koomen & Jak 2012), but some studies found a beneficial effect for students from having the same-sex teacher (Ammermueller & Dolton 2006). Some studies found that differences in ability among boys and girls seem to be independent of teachers' gender (de Zeeuw et al. 2014). Some other studies, instead, found an effect of teacher gender on recommended school type for either boys or girls (Puhani 2018). The bigger problem arises when analysing teachers' social origin, since teachers have always been considered as a homogeneous group. If we accept the idea that a match between teacher and student characteristics is beneficial for students, there are obvious reason to think that the more disadvantaged students will be penalized in terms of educational outcomes. Therefore, teacher characteristics can be seen as a big component in determining the reproduction of inequalities in educational opportunities.

Teachers Expectations and the Pygmalion Effect

Concerning tertiary effects, teacher expectations are one of the most discussed sources of educational inequalities that concerns teachers, especially in social psychology research. Rosenthal and Jacobson (1968), and then Brophy and Good (1974), demonstrated how biased teacher expectations can induce students to perform in class in a way that is consistent with those expectations. Cooper and Tom (1984) managed to define three types of teacher expectations. The first type, *estimates of present ability,* concerns the present evaluation of students' competencies in specific domains. The *expected improvement* involves teachers' prediction about whether, and how much, a student can improve his/her ability over a period of time. Finally, the *discrepancy between teachers and tests* implies a positive or negative difference between the teacher evaluation of a performance and standardized test scores (or objective ability assessment).

Teachers, on the basis of their expectations of their students, differentiate their behaviour towards different students (Babad 2009, Rubie-Davies 2018), and this has consequences on students' perception of their ability. Teacher expectations can have an effect on students in two different ways. The discrepancy between teachers and tests and the expected improvement can result in the so-called "self-fulfilling prophecies". Merton defined the self-fulfilling prophecy as a "definition of the situation [which] evokes a new behaviour which makes the original false conception come true" (1957:423). Self-fulfilling prophecies are defined in literature also as "self-maintaining expectations" (Babad, Inbar & Rosenthal 1982; Brophy 1983). When teachers have specific expectations about the intellectual ability and growth of a student, the latter is more likely to actually show this intellectual ability. On the contrary, when teachers do not have such expectations, a good performance and the intellectual growth are discouraged, and students do not perform as well as they could. This is called Pygmalion phenomenon (Rosenthal & Jacobson 1968). Self-fulfilling prophecy can be conceived as an outcome of "labelling" students as high or low performers. In this regard, there are three main causal paths by which teacher's expectations contribute to shape differences in children's achievement through:

> "First, teachers may provide higher quality instruction to students from whom they expect more. Children from groups who are the beneficiaries of higher expectations will benefit from greater exposure to high-quality instruction. Second, students may perceive cues about what the teacher expects, internalize the expectation, and become motivated and achieve consistent with the perceived expectation (Brophy & Good 1970, Darley & Fazio 1980, Weinstein & Middlestadt 1979). Third, children from academically stereotyped […] groups may, in the face of a low teacher expectation, become concerned about being judged on the basis of the stereotype, increasing susceptibility to negative expectancy effects (McKown & Weinstein 2002 2003, Steele 1997, Steele & Aronson 1995)" (McKown & Weinstein 2008:236).

The second effect derived from teacher expectations are the "sustaining expectation effects", that occur when the performance is sustained over time (because of expectations) rather than changes for other reasons than the teacher (Cooper & Good 1983).

After almost 50 years from the publication of the study of Rosenthal and Jacobson, an extensive research literature has been produced in order to explain how teacher expectations are created and how they can impact student educational outcomes. Some of these studies show that student achievement is predicted by teacher expectations because the latter are accurate rather than because they produce self-fulfilling prophecies (Brophy 1983, Jussim 1989, Thys 2018). Other studies demonstrated how teacher expectations predict accurately also students' behavioural characteristics (Van Houtte et al. 2013). Some authors investigated how students' characteristics may shape teachers' expectations (see Jussim & Harber 2005 for an overview). Given that the determinants of student outcomes are multiple, complex and intricate, overall, this research demonstrates that "teacher expectations do play a role in how well and how much students learn" (Cooper & Tom 1984:77).

Teacher Stereotypes

After having shown that teacher expectations can influence students' ability and competences, now the goal is explaining how some dimensions related to teachers' "perceptional biases" (see Jussim & Harber 2005 for a review) may impact the way in which teachers can form specific expectations and therefore evaluate their students, despite their actual competences and abilities. The main streams of research on this topic stresses the potential effect that the abovementioned expectation bias may have on students' performance and broader on students' attitudes, mental health and self-confidence (as recent examples, see Zhu et al. 2018; Boerma, Mol & Jolles 2016; Urhahne 2015; Gilbert et al. 2014). However, a simultaneous stream of research focused on finding associational relationships between students' characteristics and teacher perception bias (as recent examples, see Hornstra et al. 2018; Zhu et al. 2018; Riegle-Crumb & Humphries 2012). According to this literature, teacher expectations are strictly linked to stereotypes. Indeed, teachers' expectations are based on two different sources of information: the first one is the interaction between student and teacher, and the second one is information that come from other sources (Rist 1977). An example of an additional source of information is stereotypes.

Stereotypes, as representations of characteristics of specific groups (Bordalo et al. 2016), are means used by teacher in order to process information about students in an easy, fast and efficient way. However, they may result in biased judgement and in over or under-evaluation of specific performances. They may also lead to discrimination against specific groups, which can in turn lead to self-fulfilling prophecies by influencing the behaviour of the group (Alesina et al. 2018). Stereotypes arise in relation to the context in which the evaluation occurs, and similar achievement may be assessed in different way because the average evaluation varies between groups of students with different background characteristics; consequently, being assessed in a certain way depends on "suitable properties" of the group acting in the specific context (Correll & Benard 2006). Indeed, the stereotypical influence of student background characteristic on teacher expectations suggests that even if teacher judgment may be somehow accurate, they are inevitably bounded by prejudice (Van Houtte et al. 2013).

The main sources generating stereotypes are ethnicity, gender and socioeconomic status of individuals. In this particular context, student gender, ethnicity and social origin can shape teacher expectations according to the stereotypes linked to these groups. For example, if the context in which the evaluation occurs is traditionally associated with male characteristics (e.g., STEM fields), male students are likely to be highly evaluated respect to female students (Ridgeway & Smith-Lovin 1999). Overall, teacher expectations are lower for low SES students and for students belonging to ethnic minorities (Rubie-Davies 2006; Tenenbaum & Ruck 2007; van den Bergh et al. 2010; Glock & Krolak-Schwerdt 2014; Sprietsma 2013; Tobisch & Dresel 2017). The higher expectations are therefore for students with high social background and non-immigrant background. Jussim and colleagues (1996) explained how understanding teachers expectations linked to stereotypes is necessary in order to grasp the role of teachers in educational inequality.

Grading Bias

The empirically measurable consequence of teacher stereotypes linked to specific groups of students may lead to what are called *grading bias*. Grading bias, or gaps in grading, occur when "a teacher gives students of different [groups] grades that systematically differ but not due to their performance; this can be caused by numerous factors including intentional or unintentional discrimination" (Protivínský & Münich 2018:141). In the economics of education literature, there are two different types of gaps. The first one is the gap in ability, therefore differences among groups in their actual competences. The second type of gap is the one in educational achievement, therefore in grades, in contrast to test scores. With "grading bias" we refer to the second type of gap. A great challenge in educational research is disentangle discriminatory behaviours to non-discriminatory behaviour associated to the same outcome, for example teacher evaluation of students.

Grading bias are therefore assessed identifying systematic differences in students' assessment between blinded and non-blinded tests, according to groups of students with different background characteristics. Indeed, blinded assessment implies that information about the student identity is hidden, therefore differences between students' test scores are merely due to differences in abilities. Usually, blinded assessments are nationally measured through standardized test score, that are assumed to capture students' subject-specific competences. Grading bias are not necessary in line with teachers' expectations. From one side, teachers may be likely to assess students according to their expectations and to their "cognitive frame". From the other side, they may be conscious of the existence of such bias and therefore compensate unprivileged students by helping them with less hard standard of grading (Behaghel et al. 2015). Therefore, a compensation effect, or "reverse" discrimination, may happen, primarily in contexts and countries in which equality of educational opportunities is considered extremely important, especially if we acknowledge that evidences in the modest body of literature that compares blind and non-blind assessment underlines that grading bias are not negligible (Brennan 2008).

Gender grading bias has been widely demonstrated in a number of studies. Expectations related to students' gender are explained through the stereotypes linked to the cultural belief that men and women are "innately and fundamentally different in skills and interests" (Riegle-Crumb & Humphries 2012:291) and therefore have different capacities of dealing with different subjects. The stereotypical impression that men are better than woman in scientific fields is widespread in most societies (Bordalo et al. 2016). It is shared by most individuals, and this includes also teachers (Carlana 2018). This phenomenon regards mainly the traditional scientific-humanistic divide – that has more recently shifted in the care–technical divide (Barone 2011). Despite that, a number of studies have shown that teachers assess girls' performance higher than boys' one (Lindahl 2007, Emanuelsson & Fischbein 1986). Several theories aim at explaining such differences in grading. "Teacher-student interaction" theory (Mechtenberg 2009) explains how different gender gaps are the result of teacher and students' behaviour in class. Other studies focused on if and how students may benefit from having the same-sex teacher (Spilt, Koomen & Jak 2012; Ammermueller & Dolton 2006), and it has been suggested that the increase in the share of female teachers - the so-called "feminization of schooling" - may explain the gender-gap in achievement in favour to female. It may be that girls tend to perform better with female teachers and/or receive better grades, or vice versa, girls have benefited from the increased shared of female teachers that has been occurred in Western societies (Dee 2007). It seems that there are systematic differences in the way teachers grade female and male students with virtually the same level of competence. This is demonstrated through the comparison of teacher evaluation of a specific performance and the relative standardized test scores, blinded evaluated. In Israel, Lavy (2008) found that girls obtain higher grades than boys in "non-blind" tests in which the evaluator did know the identity of the student compared to "blind" test. Likewise, Lindahl (2007) observed that, when comparing students with the same level of competence, teachers assessed girls more generously than boys in Sweden, as Angelo (2014) reported for Portuguese high schools. Enzi (2015) found that gender plays a role in grading in Germany upper secondary education, a result that echoes that of Kiss (2013) in lower educational levels. Furthermore, also among 15-years-old Czech students there is a sizeable gender gap in teacher gradings in favor of female students, both in language and mathematics (Protivínský & Münich 2018).

Teachers' expectations, or "perceptual biases" (Jussim 1989), linked to students' SES can be nested in the theoretical framework of the cultural reproduction theory (Bourdieu & Passeron 1990). Specifically, Bourdieu's works on the different "forms of capitals" suggest how dominant groups in society appropriate some resources for maintaining processed of social (and educational) reproduction (Donnelly 2018). Cultural capital, as "subtle modalities in the relationship to culture and language" (Bourdieu 1977: 82) is detectable as linguistic preferences and styles, ways of speech and move, tastes, mannerism. The schooling context embodied the cultural capital of dominant groups in society. Those children whose parents taught certain ways of being and disposition, coherent with the dominant groups, are those more likely to be evaluated positively, because they embodied the expected culture of the educational context, and match what teachers expect from them. Cultural capital translates into educational performance though Bourdieu's (1990) concept of "institutional habitus" that describes all the dispositions, styles and behaviours which predispose individuals and groups to think and act in particular ways in specific institution (Donnelly 2018). Some kind of bias in the educational system leads to (mis)conceive cultural capital as academic ability and capacity, in such a way that students with a higher stock of cultural capital inherited from their parents may give an impression of brilliance successively rewarded in terms of grades by their teachers (Cole & Mendick 2006).

Similarly to Bourdieu's concept of "habitus", also Bernstein's work (1975, 1996, 2000) may offer an alternative approach for analyzing how educational institutions mediate the relationship between socioeconomic background and education, focusing on what happens within the school context. Bernstein's conceptual framework captures the relationship between dominant and subordinate social groups through education, considering two interrelated practices within schools: the “expressive” order, concerning characters and manners of the students, and the “instrumental” order, concerning aspects of the school that relate to the acquisition of specific skills and knowledge.

The higher stock of cultural capital, visible through what is considered appropriate conduct, character, and manner, can also lead students to show more self-control and engagement in the classroom, which has been demonstrated to induce teachers to give them better grades comparing students with the same standardized test scores (Cornwell et al. 2013). Teachers may also be likely to reward students who resemble themselves, and this “like me effect” (Fleming 1999) may be a source of biased marking, which in turn may lead to social differences in grading. Some empirical studies confirm the tendency of teachers to over-evaluate students with higher social background and under-evaluate children from less advantaged families (Helland 2007; Bygren 2020).

Students’ ethnic background can affect to a large extent teachers’ expectation. Dee (2005) finds that teachers tend to evaluate the behaviour of students with a different ethnicity than their own as more disruptive, inattentive, and more likely to not be able to complete their tasks. Generally, ethnic minority students on average perform poorer, therefore teachers may expect the latter to perform lower. However, there are several ways in which teachers react to these expectations. Teachers may give minority students higher grades if they want to encourage them, or if their performance overcomes teacher expectation. At the same time, teachers may give minority students lower grades if expectations do not allow them to recognize the real performance, or if they have a negative attitude toward ethnic minorities group in general (van Ewijk 2011). One important factor that impact the evaluation of students belonging to ethnic minorities is the fact that they are usually taught by teachers belonging to the ethnic majority. Again, the “stereotype threats” theory (Steel 1997) underlines how cultural dissimilarity between students’ and teachers’ ethnicity may favour a biased evaluation of student ability.

Regarding grading bias according to ethnicity, results are somehow contradictory. In some cases, there were no differences in grading between native and non-native students (Van Ewijk 2011; Hinton & Higson 2017). In other cases, instead, foreign students appear to have lower grades compared to native students when their names were disclosed (Hinnerich et al. 2015; Sprietsma 2013). Some other studies found a bias in grading practices according to students’ ethnicity when comparing students’ assessments and their deviation from their ability measured by national standardized tests (Ouazad 2008; Lindahl 2007). Alesina and colleagues (2018) demonstrated how the bias is driven by the subject: math teachers are more likely to give lower grades to immigrant students compared to native students when they have the same score on standardized tests, while concerning literature teachers, it seems that stereotypes do not affect the assigned grades. Actually, stereotypes seem to have a positive effect for first-generation immigrants, which are the least familiar with the language and therefore need more help. In this case, even if the stereotype against immigrant exists, teachers tend to have a positive grading bias that may help disadvantaged students to cope with difficulties linked to their non-native status. An interesting work by Burgess and Greaves (2013) found that the direction of grading bias may depend also on students’ ethnic group: if the student had an Asian background, he/she was systematically over-assessed, especially in the STEM subjects, while black and African students were systematically under-assessed.

Grading Practices

One of the features that describes teacher effectiveness is the proper grading practices, but they also may represent another way by which teachers assess differently their students. For example, a teacher with high grading standards tends to give good grades only to very high achievement, or in other word to students who show very high levels of ability (Bonesrønning 2004). Previous research about grading practices starts from the student-teacher interaction model proposed by Correa and Gruver (1987). They conceptualize student achievement as a function of student and teacher effort, where teacher grade becomes an instrument that helps teachers avoiding students' substitution of their effort for teacher effort. In the utility-function of students, grades are thought as the product between student actual ability and teacher grading practices. But teachers can intervene in the relationship between actual ability and perceived achievement with grading practices, for example emphasizing the effort-component when giving grades.

Grading practices have been demonstrated to affect student ability and knowledge on average later on (Betts 1995, Bonesrønning 2004). This is because when teachers have high grading standards, students need to increase their effort in studying if they want to achieve a good grade. However, previous studies are sometimes contradictory. High grading standards can be more effective for already high achievers, because they are motivated to increase their effort, and at the same time they may have a detrimental effect on less able students who tend to give up when they perceived standards are impossible to reach (Betts & Grogger 2003). Betts & Grogger (2003) analyzed the impact of grading standards on different outcomes and considering also students' ethnicity. Since high grading standards appear to have noticeable effects on the ability (measured through test score) of students who are already in higher position of achievement distribution, it seems that they may help increasing inequality among students. However, this does not mean that high grading standards have a negative effect on students collocated in lower position of achievement distribution: indeed, they have still a positive effect on their abilities, but tinier (Betts & Grogger 2003). The same happens when considering black students, for which the positive impact of high grading standards is smaller. Concerning high school and college attendance, high grading standards have been demonstrated to have no significant effect for white students, but interestingly they seem to have a negative effect on graduation rates for black and Hispanic students (Betts & Grogger 2003). These results are consistent with the *relative performance hypothesis* (Loury & Garman 1995), according to which students value their academic success not in absolute terms, but comparing themselves to their classmates: students at the upper bound of the ability distribution are gaining more for high grading standards, therefore the gap between lower and upper bound is higher, and consequently students at the lower bound may feel discouraged and they may give up.

When considering grading practices, one important dimension is student motivation and effort. Grading practices could be a useful tool by which student effort and motivation can be manipulated, and they could be used as a strategic mean for affecting the student’s effort and therefore achievement (Iacus & Porro 2008). In this regard, some studies focused on how the students’ effort respond to being graded and ranked (Levitt et al. 2012; Jalava, Joenses & Pellas 2015). However, effort is triggered differently according to students’ characteristics. For example, while generally students tend to exert low effort in standardized tests, boys and students with high SES show a larger difference in achievement between low and high-stake tests than females and students with low SES. Boys are also more responsive than girls to short-term incentives, while girls are more intrinsically motivated. Therefore, teachers' decisions about the proper grading practice they should adopt is not an easy task, and identical grading practices impact in a different way student with different characteristics. Moreover, teacher grading practices can affect both students' academic performance but also students' perception of their ability.

Teacher Recommendations

The second path through which tertiary effects are reproduces regards teacher recommendations. After an educational transition point, students have to choose typically between remaining in school or entering the labor market, and if they choose to remain, they can opt for an academic track or for a technical or vocational track (Breen & Goldthorpe 1997). According to Boudon (1974) students with similar competences and similar academic performance but with different economic background are likely to make different choices regard their educational future. Indeed, children and parents with higher SES are more likely to choose academic track, which are more demanding but also more remunerative later on (Boone & Van Houtte 2012; Jæger 2009). Also, ethnic background has been demonstrated to influence school choices, since non-native students tend to have more ambitious aspiration than their native counterparts (Jackson, Jonsson & Rudolphi 2012; Teney, Devleeshouwer & Hanquinet 2013). Regarding gender, female students are more likely to choose less technical and scientific tracks compared to male students, even when the academic achievement is similar across subjects (for example, see Ceci & Williams 2007).

In addition to all the mechanisms related to secondary effects that can help explaining these differences in school decisions among students with different characteristics but similar academic performance (see Jackson 2013 for a review), tertiary effects play an important role. Indeed, tertiary effects can be described as the variation in teachers' attitudes towards children of different social origin, ethnicity and gender and how this variation in attitudes have consequences for the allocation of students in specific tracks.

Given the role of teacher expectations in influencing students' ability and students' self-perceived competence and academic performance, it is reasonable to assume that teacher expectations have an effect also on educational decisions. Thys (2018) explain how this is likely to happen in two ways. First, teachers' higher expectations lead students and parents to feel encouraged in choosing more demanding academic options. Second, teachers themselves may be guided by their expectations in recommending specific tracks to students.

The fact that teachers have an influence on students' and families' decisions is supported by several studies. The subjective evaluation of the probability of a student to succeed in a chosen educational path is predicted also by teacher expectations. Becker (2013) suggests that self-fulfilling prophecies have an indirect influence on educational decisions through their impact on students' abilities and therefore academic achievement, since according to Breen and Goldthorpe (1997), perceived probabilities of success depend on student perceived ability. Indeed, students' aspirations are higher when the teacher has higher educational expectations, even controlling for student achievement (Frost 2007; Buyn et al. 2012; Thys 2018).

The second way in which teachers influence educational decisions is linked to teacher recommendations. Teacher recommendations about students' future educational path is a huge component of students and parents' school and track choices. It is an official advice, therefore institutionalised, and, in some context, it can be binding for accessing the next educational step. Depending on countries regulation, teacher recommendations may be the only "official" advice or they may be accompanied by standardized tests, and they may occur in just one transition point or in more transition points of students' educational career. Some studies show how the allocation of students into tracks is reproduced by teachers and schools because teacher recommendations are biased according to students' economic background (Barg 2012; Boone & Van Houtte 2013), ethnicity (Bonizzoni et al. 2016) and gender (Carlana 2018). Specifically, teachers are less likely to recommend academic tracks to students with an ethnic minority background or with less economically advantaged family. Gender stereotypes also influence school choices, so that teachers are more likely to induce girls to attend vocational schools to a higher extent. In addition, the stereotypical impression that men are better than woman in scientific fields, which is widespread in most societies (Bordalo et al. 2016), may affect also the allocation of students according to the traditional humanistic-scientific divide – that has more recently shifted in the care-technical divide (Barone 2011).

Teacher expectations lead to biases in teacher recommendation. However, Bonizzoni and colleagues (2016) demonstrate that concerning ethnic bias in recommendations, teachers consciously recommend less ambitious tracks to minority students because they want to protect them from failure, since they lack of cultural and linguistic resources that are thought to be essential for succeed in academic tracks. An inverse pattern occurs considering students with average performance – who can possibly succeed in academic tracks but are suitable also for lower tracks – but whose parents are highly educated. Argentin et al. (2017) show that when students have similar school proficiency, teachers are more likely to suggest the academic tracks to those students whose parents are highly educated. This has consequences in terms of probability of success for students with average academic performance and higher socio-economic background who decide to enrol in the academic track, because the risk of failure is significantly higher.

Teachers can be very influential regarding differentiations in students' treatment, which can be more or less favourable due to varying student social status, cultural capital, ethnicity or language codes, but they are very influential also concerning the process of allocating students to different school tracks (Reimer 2019). There are other mechanisms identified by scholar through which teachers could be driven to orientate students with specific characteristics to better schools, even when students have equal academic performances (Argentin et al. 2017). Those mechanisms are: evaluation of students' non-cognitive skills and of pupils' attitudes (Timmersmans et al. 2016); perception of parental support in school subjects (Barg 2012; Mayer et al. 2015); differential in educational and occupational aspirations of parents (Stockè 2007); direct favouritism for highly educated parents and children (Farkas et al. 1990).

The Educational Context: Schools and Classrooms

The core of this chapter is the role of teachers and how they may impact the accumulation or the compensation of inequalities in education. However, teachers operate every day in specific settings which may shape the relationship between teacher and student, and therefore may also influence student educational outcomes. Specifically, classroom and school characteristics are substantial component of the teacher-student equation. In this final section, some features of schools and classroom that may affect teacher-student relationships are briefly introduced.

Classroom characteristics may be shaping teachers' perceptions and expectations. Mashburn and colleagues (2006) explore how some classroom characteristics are associated with teachers' perception of students' competencies, such as the child-teacher ratio and the number of hours student spend in the class per day. Results show that students in those classrooms with longer school days in terms of hours were more likely to be in a conflictual relationship with their teacher, probably because when the time spent together is longer, teacher are more exposed to more negative behaviours. In addition, a lower child-teacher ratio is associated with higher ratings of children competencies.

Concerning child-teacher ratio, there has been and open debate about the educational consequences of differences of class size (Blatchford, Bassett & Brown 2011). The attention of scholars has been focused on whether the size of the classroom can influence pupil academic outcome, but results are controversial (see Wilson 2006 for a review). However, some studies agree on the fact that small classes lead to higher achievement (Finn & Achilles 1999; Blatchford et al. 2003). In order to understand the effect of class size, it is necessary to understand what happens daily in the context of the classroom, specifically looking at the processes such as the interaction between students and the teacher, student engagement and involvement, and classroom control and management. Concerning the effect on teaching, it has been explored whether bigger classes can lead to a decrease in the amount of time that teachers can dedicate to single students and to build a solid and trustful relationship with them. However, there is no agreement on the size of the effect (Bruhwiler & Blatchfort 2011; Blatchford et al. 2002; Ehrenberg et al. 2001).

Even the share of female students in the classroom have been demonstrated to have effect on academic outcomes. Hoxby (2002) found that both male and female students tend to have higher achievement when they are in classes with larger shares of girls. This can be related to different mechanisms: these classes may have more disruptive students, less pressure to be feminine and therefore unenthusiastic feelings toward scientific fields, a more relaxed environment that may allow teachers to be more effective, and so on. Other studies focused on the classroom's structural environment and the classroom's symbolic environment. Structural factors – such as lightning, acoustic, temperature, air quality, and symbolic factors – such as classroom layout, objects and décor, on which teachers have a direct control, have been demonstrated to influence students' achievement (Cheryan et al. 2014).

Concerning schools, one of the determinant factors that has to be considered is the type of school leadership. Some quantitative research has conceptualized how leadership, that is school principals, can have an indirect effect on student outcomes through the establishment of teaching conditions (e.g., providing teacher professional learning opportunities, teacher job satisfaction) (see Marzano et al. 2005). Indeed, school leaders have an indirect effect on student outcomes which is mediated by teachers (Robinson et al. 2008; Hallinger & Heck 1998). More in general, the school quality is directly linked to the probability of grade completion. Indeed, students are less likely to remain in school when they are attending a low-quality school rather than a high-quality one, controlling for achievement (Hanushek et al. 2008). However, the quality of a school is defined mostly through the effectiveness of teachers working in that context. But cyclical phenomena occur, through which higher quality school are defined trough their teachers, and higher quality teachers are more likely to be assigned or to choose to work in higher quality school. Students coming from a privileged background have a higher probability of study in a higher-quality school and therefore of being matched with a well-performing teacher, reproducing inequalities (Abbiati et al. 2017; Goldhaber et al. 2015; Sass et al. 2012).

Another situation that can create unintended bias is the common non-random sorting of teachers and students, both across and within schools (Dee, 2005). Erhenberg and colleagues (1995) find that when minority students with low performances are systematically assigned to minority teachers, the effect of having a demographically similar teacher understate the results of grading bias. In this regard the quality of schools is also linked to the geographical collocation. Indeed, there may be great differences in the quality of schools located in different areas of the same country. For example, Abbiati, Argentin and Gerosa (2017) targeted teachers in order to explain the variability in quality that exists among schools and Italian macro-areas, exploring the allocation of teachers to students. Clearly, if there is a systematic student-teacher match on the basis of their characteristics, educational inequality is more likely to be reinforced. The results show that the more experienced and effective teachers are systematically paired with high performing students. These associational patterns have been found across different schools but also within school between classes (see also Isenberg et al. 2013), meaning that a segregation of students occur also at the school level. This is strengthened through teachers, who tend to leave low-quality schools – where the teaching conditions are harder – when they reach seniority status, in favour of high-quality schools, characterized by less problematic, disadvantaged students and higher background students. In the study of the role of teachers in shaping educational inequalities, it is therefore necessary to account also for the geographical variability in socio-economic contexts, in the management of schools and classrooms and in the functioning of the educational system (Pavolini et al. 2015). In conclusion, structural features of schools and classrooms may influence teachers' positive or negative perceptions of students' competencies, but also teaching experience, in every contradictory aspect.

Conclusions

This theoretical chapter explores the complex relationship between teachers and their students. Starting with a broad definition of tertiary effects, an explanation of the several mechanisms that come into play in the relationship between teachers and student assessment from one side, and student choices from the other side, is proposed. In this framework, teacher expectation bias can be considered as the core mechanism in shaping inequalities in educational opportunities linked to students' gender, ethnicity and socioeconomic background. Indeed, teachers' expectations are directly connected to teacher grading bias and teacher recommendation bias.

Adopting a sociological perspective can be extremely useful in order to gain more knowledge about whether and how teachers can play a role in the reproduction of educational inequalities. Dee (2005) differentiates between a *passive* teacher effect and an *active* teacher effect. Indeed, it is still not clear if teachers (un)consciously behave in ways that may help the most struggling students or if they operate according to their prejudice, widening the preexistent gap between students coming from different social groups. Moreover, there is still much to understand about differences between school-grades, fields and/or subjects. Sociological tools can help in adding more value to the already existing studies about teachers, since it can combine information at the individual level for both students and teachers with institutional and structural characteristics of classes, schools and even educational systems that can influence the way in which teachers interact with their students. Moreover, sociological tools vary from in-depth qualitative ethnography to statistical analysis on large n, and adopting different approaches may help shed a light on this intricate phenomenon

CHAPTER 2

THE STRICTER THE BETTER? THE IMPACTO OF EARLY TEACHER GRADING STANDARDS ON STUDENTS' COMPETENCES DEVELOPMENT AND ACADEMIC TRACK ENROLMENT[2]

Abstract

Despite the growing attention on teachers' grading practices in educational research, less attention has been dedicated to the consequences of teachers' grading standards on students' educational outcomes, especially in early stages of their scholastic career. This chapter aims at filling this gap, analysing the impact of teacher's severity in grading on students' competences development and academic track enrolment, and how it varies according to students' gender, socio-economic background and immigrant status. The analysis relies on Italian INVALSI-SNV data: information on 5th grade students and their teachers are linked, and pupils are then followed up to 8th and 10th grade, in which their competences and school track are recorded. Relying on 2SLS regressions, findings show that being exposed to stricter grading in 5th grade leads to higher students' competences later on, and to higher probability to enroll in traditional lyceums, with no notable heterogeneous effects across students with different characteristics.

Keywords: teachers, grading standards, academic outcomes, student competences, school track.

[2] Chapter 2 is co-authored with Moris Triventi and Emanuele Fedeli, and submitted in *Social Science Research.*

Introduction

The analysis of grading practices, that is the way in which teachers grade their students, is at the core of an extensive literature in educational studies (for a review on teacher judgments, see Urhahne & Wijnia 2021). Grading practices have been shown to have a substantial impact on students' educational outcomes. Existing studies on this topic agree on the importance of grades in contemporary educational systems as well as in the labor market (Tyner & Gershenson 2020). Teacher grades can affect students' learning processes and how they perceive themselves in terms of ability and competence, which in turn have long-run implications for several students' life outcomes.

Among the immediate consequences of grading practices, there are students' placement in classroom, grade promotion and attendance habits (Bonner & Chen 2019; Gershenson 2016). Medium and long-run consequences might involve students' school choices, occupational decisions and earnings in adulthood (Borghans et al. 2016; Chetty et al. 2014; Bonner & Chen 2019).

Among educational institutions worldwide, grades serve as fundamental sorting and signaling mechanisms (Chowdhury 2018). However, these signals are not provided only to the students, who may need them in order to form an idea about their intellectual ability and, consequently, their possible educational future, but are captured and reproduced by many other players in the educational arena. With the increasing complexity of the educational systems, the significance of grades has assumed numerous facets, as many as the actors involved such as parents, teachers, principals, colleges and firms. For example, grades are important signals allowing a communication of students' academic achievement between schools and families. Parents use teachers' evaluations to make educational choices for their children and to efficiently track them in the school system, and to understand if their children need educational support (Jalava et al. 2015).

Correa and Gruver (1987) conceptualize grades as a fundamental parameter in the students' utility function. Since students care about teacher's perception of their achievement, students' effort and achievement may be affected by how teacher decide to grade students (Iacus & Porro 2008). Indeed, teachers can decide to adopt certain grading standards, that is the ability level needed by students in order to get a specific grade, or, in other words, how stringently teachers assess their students. Teachers with higher/harder grading standards tend to give good grades only to very high achieving students, who show very high competences and ability levels, while teachers with lower grading standards are likely to give good grades also to those students with average levels of ability, shrinking the grading scale.

Despite the large public debate on teachers' adoption and implementation of specific grading practices, especially in primary education, little empirical research focuses on how teacher can influence students' effort and motivation adopting specific grading standards, and on the associated educational consequences. On one side, students whose teacher adopts higher grading standards are those who need to put more effort and to study more if they want to achieve a good grade, and as a consequence, students might benefit in terms of competences in the long run (Iacus & Porro 2008). On the other side, higher grading standards may discourage students if the level of ability needed for achieving a good grade is too high. Moreover, it has been hypothesized that high grading standards may have heterogeneous effects among students (Betts & Grogger 2003) since motivation may be triggered differently according to students' characteristics (Becker & Rosen 1992) such as gender, socio-economic background and immigrant status.

The aim of this chapter is to contribute to the empirical research on the effect of grading standards adopted in primary schools on educational outcomes, relying on a causal approach. The goal is to understand the effect of teacher grading standards measured in 5th grade on students' competences in 8th and 10th grade in two subjects – Language and Mathematics, and on school track in 10th grade. Additionally, the aim is to analyze whether teacher grading standards may have heterogeneous effects according to students' sociodemographic characteristics such as gender, socioeconomic background and migratory background.

The focus is on the Italian educational system, which is well suited for the study of teacher grading standards and their consequences, because teachers have a great deal of autonomy and independence in deciding their own grading practices, also when considering the school administration (Bracci 2009). On the one hand, this allows a certain degree of variation in grading practices already in early stages of educational career. On the other hand, Italian teachers' autonomy in deciding grading practices permits to explore the consequences of grading standards measured at the classroom level instead of at the school level, leading to more fine-grained results.

I rely on the INVALSI-SNV data, focusing on a cohort of Italian 5th grade students in the academic year 2013-14, which is followed up to 8th grade (a.y. 2016-17) and to 10th grade (a.y. 2018-19). This dataset allows to match students with their teachers and therefore to control for students', teachers' and classrooms' characteristics. Moreover, the availability of both teacher grades and students results in standardized tests allows to create a measure of teacher grading practices.

The contribution of this article to the understudied topic of grading standards are threefold. First, as abovementioned, the Italian data used permits to explore the consequences of grading standards at the classroom level, instead that at the school level, thus providing a more realistic and fine-grained perspective. Second, very few empirical research investigates the impact of more rigorous grading standards measured at the early stages of educational career (see Figlio & Lucas 2004 for an example), when there may be stronger effects on children self-perception of their ability, motivation and future educational choices (Facchinello 2020). Third, this contribution allows to causally assess the long-term consequences of having a teacher with high/low grading standards, not only in terms of competences but also in terms of academic track enrollment, which is a key transition point in the educational system associated with higher changes of later educational and labor market success (Barone et al. 2021; Triventi et al. 2021). This approach, combined with the analysis of heterogeneous effects on students with different characteristics, may have important implications also in terms of policy making, as discussed in the conclusions.

Literature Review on Grading Standards

In education, teachers' grading standards reflect the ability level needed by students in order to get a given grade. A teacher or a school with high grading standards tends to give good grades only to very high achievement, or to students who demonstrate very high levels of ability (Bonesrønning 2004). When measuring teacher grading standards with observational data (administrative or survey data), two pieces of information are usually needed at a classroom or school level: first, student ability measured by standardized test scores; second, student achievement measured by teacher assessment. The difference between these two variables provides an idea about how stringently teachers assess their students compared to their actual competences.

Previous research about grading practices is rooted in the student-teacher interaction model proposed by Correa and Gruver (1987). In their utility-function of students, grades are thought as the product between student actual ability and teacher grading practices. But teachers can intervene in the relationship between students' competencies and assessed achievement through their grading practices, for example emphasizing the effort-component when formulating their judgements. In other words, through the practice of grading, teachers can both evaluate the sheer quality of students' work and at the same time motivate and encourage them to study (Walvoord & Anderson 1998). Grades, as the most common type of feedback provided by teachers to students, can be powerful moderators of learning, but their effect is always difficult to capture. Hattie (2012) formalized the feedback function, where grades can "provide cues that capture a person's attention and helps him or her to focus on succeeding with the task; it can direct attention towards the processes needed to accomplish the task; it can provide information about ideas that have been misunderstood; and it can be motivational so that students invest more effort or skill in the task" (Hattie 2012:115). Feedbacks may work at four different levels – task, process, self-regulation and self – and may serve to challenge students, helping them setting their own goal and stimulating commitment.

Following this idea, many empirical studies that proposed and supported the idea that grades – and the practice of grading – can have a significant impact on students' academic outcomes focused mostly on higher education and on college courses choice (Clark 1969; Gold et al. 1971; Hales et al. 1971; Cherry & Ellis 2005). However, relatively few authors have focused on the consequences of grading standards in early stages of the educational career. Concerning students' competences, Betts (1997; 1998) hypothesizes that more stringent grading standards will increase effort among students, and therefore their subsequent achievement. The author focuses on 7th and 10th grade students, and findings suggest that grading standards are important determinants of high school students' competences. Betts and Grogger (2003), analyzing 1,000 high schools, also find a positive effect of harder grading standards on students' performance in 12th grade, especially in the upper end of the grade's distribution. Figlio and Lucas (2004) analyze the teacher-level grading standards on elementary students' achievement in Florida, using data on 3rd, 4th and 5th grade. They find that higher grading standards seem to benefit students in language and mathematics test scores over time. On the contrary, Montmarquette and Mahseredjian (1989) analyze the effect of hard grading – grades set below the real achievement – on Canadian primary school pupils and found that they have a negative effect on test scores in Language, while they have no effect on test score in Mathematics. Some studies on Norway find that lower secondary school students who are exposed to harder grading standards perform better in mathematics (Bonesrønning 1999; 2004). The same results are confirmed by the study conducted by Iacus & Porro (2008) on a local sample of 20 lower secondary schools in Lombardy, an Italian region, in three subjects (language, science and mathematics). Concerning the impact of grading standards on students' educational choices at earlier educational stages, empirical research is even scarcer. To the best of the author's knowledge, only Betts and Grogger (2003) showed that harder grading standards have no significant effect on high school decision or college admission in the United States for the period 1989 to 1991.

The basic mechanism behind the effect of grading standards is thought to be related to their influence on students' motivation and effort (Iacus & Porro 2008). In this regard, some studies focused on how students' effort respond to being graded and ranked (Levitt et al. 2012; Jalava et al. 2015). On one side, setting higher grading standards may induce students to study more and to put more effort in order to satisfy the requirements imposed by their teachers. Indeed, when teachers have high grading standards, students need to increase their effort in studying if they want to achieve a good grade. On the other side, standards that are too high to reach can induce students to give up, and therefore they may have a detrimental effect on their competences, making the relationship between the two possibly non-linear. Facchinello (2020) found that even being graded instead of not-being graded in the early stages of schooling have negative effects in effort among low-ability and low-SES students, who show lower motivation also later on.

It must be underlined that effort and motivation may be triggered differently according to students' ability and classroom composition. High grading standards can be more effective for already high achievers, because they have the cognitive resources to meet such high standards, but at the same time they may have a detrimental effect on less able students who tend to give up when they perceived standards are impossible to reach (Betts & Grogger 2003). Since high grading standards appear to have noticeable effects on students' competences who are already in higher position of achievement distribution, they may exacerbate achievement dispersion among students. However, rather than being detrimental for low-achieving students, higher grading standards may also translate in a smaller but still positive effect on their subsequent academic performance (Betts & Grogger 2003). The composition of the classroom can also act as a moderator of the impact of grading standards: indeed, in the United States high standards appear to be beneficial for high-achieving students when they are in low-achieving classes and for low-achieving students in high-achieving classes (Figlio & Lucas 2004).

It is important to note that given that academic performance is related to students' socio-demographic characteristics such as gender, ethnic background and social origin

(Hattie 2008), more rigorous grading standards can also affect social inequalities in student achievement. Moreover, the response to grading incentives of different groups of students may not be uniform (Chulkov 2006). Yet, there is little evidence showing how students with different sociodemographic characteristics respond to the same grading standards. Concerning students' gender, Fallan and Opstad (2012), analyzing a sample of business school students, found that male students are more responsive to harder grading practices, and they are more willing to put more effort if a change in grading standards requires more work in order to get an expected grade, while female students are less sensitive to change in grading standards. Boys are also more responsive than girls to short-term incentives, while girls are more intrinsically motivated (Vecchione et al. 2014). Motivation incentives may also trigger differently students with different socioeconomic and migratory background. For example, a recent paper investigating a sample of Italian students demonstrates that lower socioeconomic background is associated with lower level of intrinsic motivation and higher level of amotivation (Manganelli et al. 2021), and this may be associated with a more positive response to harder grading standards considering grades are a tangible, short-term reward. Other studies found that ethnic minority students show higher intrinsic motivation than native students, possibly to face their stigma awareness (Eccles et al. 2006; Gillen-O' Neel et al. 2011), therefore they may be less responsive to harder grading standards as a tool for manipulating effort. However, this pattern is not corroborated by a study on the Italian case, where children with immigrant parents display instead higher levels of extrinsic motivation than natives (Triventi 2020).

The Italian Grading System

In primary and secondary Italian schools, the Italian Ministry of Education (MIUR - *Ministero dell'Istruzione e del Merito*) offers indications about how teachers are supposed to grade their students. Teacher grades are assigned on a scale that goes from 1 to 10, where 6 is considered as the passing grade[3]. There are mainly two moments in which students and families meet with teachers and schools in order to know about the children's academic situation, and they correspond to two report cards. The first one is around February (first semester) and the second and definitive for that academic year is around June (second semester). If students report a grade below 6 in any subject at the end of the school year, they have to take an exam in that subject before the beginning of the new school year in September. If the result of such exam (*esame di riparazione*) is still insufficient, the student has to repeat the previous grade. Moreover, if students have three or more subjects with a grade below 6 in the final school report, they have to repeat the year, depending on the judgments of all professors for that students who join in order to decide case by case (*consiglio di classe*).

The report card usually shows average grades for each subject of all the examination undertaken by students until the end of the semester. The type of exams depends on the subjects, on the school regulation, but mostly on professors, who have a great deal of autonomy in deciding the exam structure (e.g., multiple choice questions vs open ended questions, oral exams vs written exams), the frequency for the evaluations as

[3] This is true considering the academic years under examination in this chapter. However, a recent reform in Italy (2021) has introduced a new grading system in primary schools, that consists in eliminating numerical grades in favour of more descriptive students' evaluation. This evaluation should reflect four levels of learning, approximately defined as "advanced", "intermediate", "basic" and "in the process of acquisition".

well as the grading criteria. Even if the MIUR offers some guidelines about grading practices, it is not known or clear the extent to which schools and teachers follow such guidelines: teachers usually decide their own grading criteria and grading practices, mostly according to each school's specific regulations.

After 8th grade, Italian students make their first educational choice concerning high schools. Interestingly, neither teacher grades nor teacher recommendations are binding for entering specific tracks, and formally there are no access criteria. High school can be broadly divided in vocational schools (*istituti professionali*), technical schools (*istituti tecnici*) and lyceums (*licei*). Lyceums represent the academic track, and they can be further divided in traditional lyceums and other lyceums. Traditional lyceum includes the classical lyceum, focusing on humanities, and the scientific lyceum, focusing on math and science. Generally, this is considered the most prestigious and demanding track, that leads to university enrollment. Other lyceums are considered less prestigious, and include linguistic, socio-pedagogical and artistic lyceums. Technical and vocational schools, instead, usually lead to entering the job market. Despite Italian upper secondary education is strongly stratified, university enrollment is formally open, and it does not depend on previous academic performance, or final grade: the basic requirement is having a 5-year high school diploma, although access to some universities is regulated by admission tests.

Regarding grading practices and grading standards, the topic has attracted public attention especially for what concerns the North-South divide in upper secondary education. In this respect, Argentin and Triventi (2015) examined the geographical heterogeneity in grading standards in two subjects and across the three educational levels constituting compulsory education in Italy. The results indicate that southern regions are generally characterized by lower grading standards, meaning that teachers are more generous in assigning grades for a given level of competence, especially considering high performing students. Yet, the Italian context is characterized by high levels of heterogeneity, even among provinces or schools within the macro-areas.

Research Design

Data

The empirical analysis is based on data collected by INVALSI-SNV (Italian National Institute for the Evaluation of the Education System). The main aim of INVALSI is to perform periodic, systematic and standardized assessments on students' competences. The SNV (National Evaluation System) data contain socio-demographic variables for the whole population of students enrolled in specific grades and academic years. Additionally, they contain information on both teacher assessment of student achievement (teachers' grades) and student scores in standardized tests in Language and Mathematics (INVALSI test score[4]). Both measures of teachers' grade and standardized

[4] The INVALSI (National Institute for the Evaluation of the Educational System) tests are constituted by written assessments taken by all the Italian students in specific grades (2, 5, 8, 10 and recently also 13). The goal is to evaluate, at specific key educational points, the quality of the knowledge concerning some fundamental competences in Language, Mathematics and English. INVALSI tests are equal and standardized at the national level, to allow the comparison of results between schools, municipalities, provinces and regions, as well as over time thanks to precise statistical techniques such as the "*anchoring*" of one year to another. This allows the tests to be comparable also over time and between different grades. The competences assessed through INVALSI are those required by the law concerning Italian curricula, and INVALSI elaborates the national indications for assessment (*quadro di riferimento per la valutazione*). This includes not only the evaluation of specific knowledge but mostly specific subject-specific competences, as the ability to reason about real-life issues or problems, to apply the knowledge learned, to create connections between competences and to apply them to new problems. The Italian test measures

test scores are collected from INVALSI. Starting from the year 2012, INVALSI handed out for the first time a CAWI questionnaire addressed to a random sample of Language and Mathematics teachers for specific grades. The questionnaire collects information on both teachers' socio-demographic characteristics, teaching habits and practices.

The selected sample for the analysis includes the cohort of 5th grade students in 2013/14. Leveraging the availability of unique classrooms identifiers, we matched this dataset with information from their teachers sampled in the same academic year. Students are then followed through their academic career using a student unique identifier (the SIDI code). The cohort of 5th grade students is therefore followed over time, linking

two different types of skills. The first one relates to the ability to understand authentic written documents, taken from literature, non-fiction of everyday situations. Questions deal with the nature of the text, the meaning, the intentions of the author. The second type of skills relates to the ability to reflect on the use of the language, and consequently the knowledge and use of grammar. The mathematic test measures the ability to solve problems, both within the discipline or applied in real-life situations, therefore it measures logic skills, the interpretation of graphs, the understanding of specific phenomena, the construction of models, the use in science. The number of items may vary between grades and cohort, but it is around 30-45 items for each subject which can be closed or open ended. In order to ensure the validity of the INVALSI tests, they are always pre-testing in different occasions. Moreover, results are weighted for different factors that may distort the results (school weight, class weight, cheating weight). Questions are also automatically and randomly chosen from a database of questions, which decreases the possibility of cheating, and since the chosen questions are equally difficult, the tests can be considered equivalent and comparable. The results obtained by students in the INVALSI test are measured on a quantitative Rasch scale, where 200 is the mean and the standard deviation is 40; a similar method is used by the PISA and TIMSS evaluations. INVALSI tests are an objective measure of students' knowledge. More precise information about the construction of the tests and the validity can be found at: https://invalsi-areaprove.cineca.it/index.php?get=static&pag=rapporti_invalsi.

information in 8th grade in the academic year 2016/17 and in 10th grade in the academic year 2018/19[5].

Since the language test and the mathematics test are administered in different days, some students may have been absent on one of the two days. In order to compare results across subjects, the analysis rely on a unique analytical sample that includes the considered outcomes in both the two subjects, respectively in grade 5, 8 and 10. Our final sample includes 9,370 students[6].

Measuring Teacher Grading Standards

The main independent variable is teacher grading standards, measuring how stringent teachers evaluate their students, relatively to the student achievement measured through the INVALSI test score. Standardized test scores are designed to capture specific competences acquired by students during their educational career (Heckman & Kautz 2014) and are considered more objective than grades, also because they are usually blinded evaluated. Following Betts and Grogger (2003), a measure of teacher grading standards is construct using two pieces of information: students' grades in Language and Mathematics, as a measure of how a student stands relatively to their classmates, and students' test score in language and Mathematics, as a measure of the student competences relatively to all Italian students. Teacher grading standards are estimated for each class, therefore all the students in the same class have the same teacher grading

[5] This is the unique cohort that was possible to follow, since in 2019/20 the INVALSI test was not administered due to the COVID-19 pandemic.

[6] Analysis performed with samples including the higher number of cases as possible for language and mathematics, therefore different samples for the two subjects, lead to almost identical results.

standards. Relying on two different regressions for Mathematics and Language, grading standards estimates are obtained by regressing separately students' test score in mathematics and in language competences on students' GPA (grade point average) in mathematics (eq. 1a) and in language (eq. 1b) respectively, plus a vector of classroom dummies:

$$Test\ Score\ Maths_{ic} = \beta_1 Teacher\ Grade\ Maths_{ic} + \alpha_2 Classroom\ Dummies_c + \varepsilon_{ic} \qquad (1a)$$

$$Test\ Score\ Lang_{ic} = \beta_1 Teacher\ Grade\ Lang_{ic} + \alpha_2 Classroom\ Dummies_c + \varepsilon_{ic} \qquad (1b)$$

Coefficients of classroom dummies are the estimated grading standards in Language and Mathematics. This implies that if there is a variation across teachers, a class with higher α_c has higher/harder grading standards. If $\alpha_{c1} > \alpha_{c2}$, a student in class 1 is exposed to higher grading standards respect to a student in class 2: the two students have an equal GPA in subject s, but the student in class 1 has a higher test score in subject s than the student in class 2.

Outcome Variables and Control Variables

The goal of the analysis is estimating the effect of grading standards in the 5th grade (t = 0, primary education) on students' subject-specific competences when students are in 8th grade (t = 3, lower secondary education) and in 10th grade (t = 5, upper secondary education). Moreover, the aim is assessing the effects of such grading standards in primary education on the probability of being enrolled in traditional lyceum when students are in 10th grade (t = 5). To sum up, the outcome variables are: 1) student competences in Language and Mathematics in grade 8, 2) student competences in Language and Mathematics in grade 10; and 3) the probability of being enrolled in traditional lyceum in grade 10. Competences are measured through the INVALSI test score, while the school track is retrieved from the administrative register.

The INVALSI-SVN data allows to control for a rich set of variables that concern student characteristics and demographics. I selected control variables following general recommendations from the causal graph literature (e.g., Cinelli et al. 2002). Among these, a measure of students' achievement in $t = -1$, as self-reported average grade at the end of 4th grade is included. It is reasonable to assume that this measure captures what could have influenced parents' educational choices up to $t = 0$. Regarding teachers, control variables include a set of demographics (age, gender, educational credentials, parental education) together with some indicators associated with teacher effectiveness, such as type of contract and teaching to test information. At the classroom level, control variables include share of females, share of students with high socioeconomic background, share of immigrant students and class size. For a more detailed description of the control variables, see appendix Table A1 and Table A2.

Methods

The goal of this study is to causally identify and estimate the average treatment effect of being exposed to a particular grading standard in 5th grade on students' subsequent academic competences (in language and mathematics) and their school track placement in upper secondary education. In order to do so, I developed two distinct approaches, which rely on different assumptions. In the first approach, I provide an identification of the causal effect controlling for an extensive array of individual, teachers, and classroom characteristics, and introducing school fixed effects to control for unobserved characteristics at the school level. This first approach relies on three main assumptions: 1) No reverse causality between treatment and outcomes; 2) No confounding bias (at the individual and higher levels); 3) Teachers' characteristics are good proxies of teacher proclivities. Given that treatment and outcomes are measured in distinct moments of time, and given that we control for previous academic competences, the first assumption is likely to be satisfied. To empirically support the plausibility of the second assumption, a randomization check is performed, to evaluate whether grading standard "predicts" invariant student characteristics (Pei et al. 2019)[7]. Results show appreciable as-good-as-random distribution of grading standards across students, except for students' socioeconomic status (see Appendix Table A2.2 and A2.3).

However, the third assumption according to which teachers' characteristics are good proxies of teacher proclivities might be violated. Teacher proclivities may affect grading standards due to observed and unobserved student characteristics, and they may depend also on teacher-student interactions (Aucejo et al. 2022). To control for such bias, in the second approach, I aim to account for potential remaining unobserved heterogeneity by relying on an instrumental variable design, where the instrument is the grading standards of other classrooms in the same schools. The intuition is to exploit a

[7] I test for consistency with as-good-as-random assignment of treatment in order to assess whether our treatment is randomly distributed across student categories (e.g., based on gender, ethnic origin, socioeconomic status); A low degree of selection and a rich set of controls support the plausibility of the lack of relevant omitted variable bias (see appendix Table A2.2, A2.3). I also test for consistency with as-good-as-random assignment of teachers' characteristics to classrooms, according to classroom composition (class size, mean ESCS, percentage of immigrant students, percentage of female students), see appendix Table A2.4, A2.5.

teacher peer effect[8] within the school, where teachers are likely to discuss and compare grading practices, therefore, to influence each other's grading practices. The two approaches have in common the use of school fixed effects, to control for heterogeneity of grading standards (Argentin & Triventi 2015). Their goal is estimating the total effect of grading standards, therefore post-treatment variables which might lead to a bias are not included in the regressions (Elwert & Winship 2014). The estimation strategies follow the two approaches. In the first approach, three linear OLS regressions are estimated[9], with the following general specification:

$$Outcome = \beta_0 + \beta_1 \hat{\alpha}_{c_{t0}} + \beta_2 X_{i_{t0}} + \beta_3 T_{c_{t0}} + \beta_4 Z_{c_{t0}} + \mu_{s_{t0}} + \varepsilon_{ic} \qquad (2)$$

The three outcomes are: 1) Mathematics and 2) Language competences measured three and five years after the 5th grade, when a new sorting of students in the lower and upper secondary education occurred; 3) probability of being enrolled in a traditional lyceum five years later. In the equation, $\hat{\alpha}_{c_{t0}}$ is the treatment of interest measuring teachers' grading standards; $X_{i_{t0}}$ is a vector of individual characteristics; $T_{c_{t0}}$ is a vector of teacher

[8] Reflection is not an issue as outlined by Hernán and Robins (2006). First, IV estimation does not rely on assumptions about the causal ordering between the instrument and the endogenous regressor. (Birkelund & van de Werfhorst 2022).

[9] In order to check for the nonlinearity of the relationship, analyses are performed on the same models adding a quadratic term to the treatment variable. However, Likelihood-ratio test, AIC and BIC show no differences between the linear regression and the quadratic regression when including the control variables, even when the quadratic term is statistically significant. Results for model 3 are shown in appendix Table A2.6, A2.7, A2.8 and Figures A2.1, A2.2. The linearity of the relationship may be due to the fact that grading standards are measured in primary schools, where grading standards may generally be not particularly heterogeneous and overall not particularly severe.

characteristics; $Z_{c_{t0}}$ is a vector of classroom characteristics and $\mu_{s_{t0}}$ are school fixed effects. In the second approach, the previous equations are modified by including the first stage of a 2LS estimation:

$$GS_s = \beta_0 + \mu_{s_{t0}} + \beta_1 \hat{\alpha}_{other\ class_{t0}} + \beta_2 X_{i_{t0}} + \beta_3 T_{c_{t0}} + \beta_4 Z_{c_{t0}} + \varepsilon_{ic} \qquad (3)$$

Where GS_s represents the subject-specific grading standards as estimated in equation 1, $\hat{\alpha}_{other\ class_{t0}}$stands for the grading standards adopted in the other classrooms in the school (the instrumental variable); all other terms are previously defined.

An additional empirical issue I tackled in the estimation of our statistical models refer to longitudinal missing values, commonly known as 'panel attrition'. Indeed, following students through their academic career implies an attrition that causes a significant loss of cases from the initial sample. This is due to several factors, such as grade retention, students transferring, non-reporting of SIDI codes by school administrations and potential misclassification of SIDI codes. This may lead to a possible selection of high performing students that may in turn affect the estimates. I consider the possible selectivity of students observed throughout the entire time span considered (from 5th to 10th grade), by correcting the estimates with an inverse probability weighting (IPW) approach, which has been shown to be effective a wide range of settings (Seaman & White 2013). In order to construct IPWs, I estimated a binomial logistic regression on the probability of being observed in the 10th grade among 5th grade students with valid information, as a function of a number of students' characteristics[10]. Then I computed

[10] The covariates are gender, quarter of birth, ethnic background, regularities of studies, geographical area, attendance to infant school, attendance to kindergarten, socioeconomic background, INVALSI test score in Mathematics and Language, test anxiety. In order to control

predicted probabilities based on this model, I created weights as the inverse of the predicted probability and incorporated the regression estimations.

Results

Descriptive Analysis of Grading Standards

Figure 2.1 represents the distribution of grading standards (standardized) in the two subjects. In order to understand how grading standards are interpreted, it is important to recall that, through the analysis, the measure of how stringent the teacher is when assigning grades is not interpretable in absolute terms. Indeed, the construction of GS is relative to the selected sample – therefore to the selected teachers: the estimated effect on students' educational outcomes is interpretable as a change in severity within the selected population.

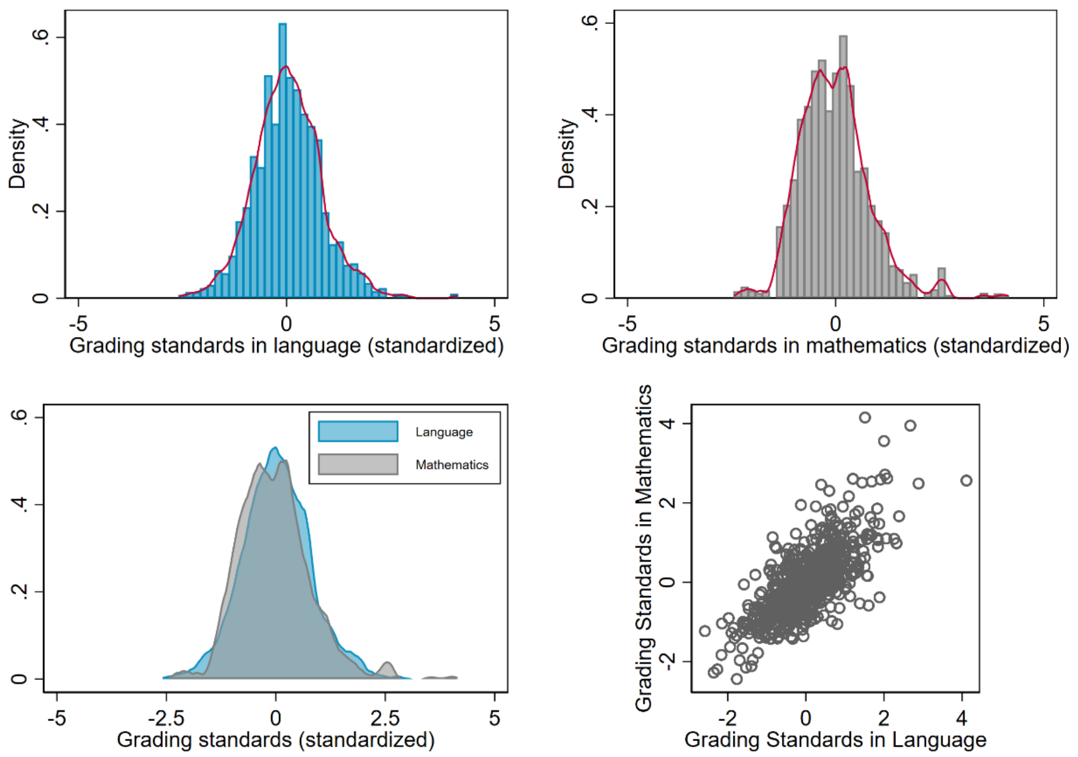
However, considering that the analysis relies on a random sample of the whole population of Italian students in 5th grade in the academic year 2013-14, it is reasonable to assume that grading standards manages to virtually capture the whole spectrum of teacher severity in the considered grade.

In order to understand how to interpret grading standards in Language and Mathematics, it may be useful to rely on Table 2.1, in which classrooms with the lowest and the highest grading standards in Language are reported and compared with the

for the validity of IPWs, we perform additional analyses with weights attributed to students as the mean of the respective quantile of IPW, and results show no significant differences.

classroom average score and the average grade. In order to facilitate the interpretation, we also report the mean for the eight values for GS, score and grade.

Figure 2.1: Distribution of grading standards in Language and Mathematics (N = 9,370) and correlation between grading standards in Language and grading standards in Mathematics



It is noticeable how classes with lower grading standards, therefore having a more generous teacher, have poor INVALSI test score results compared to classes with higher grading standards, therefore having a stricter teacher (mean score of 179 against mean score of 262). At the same time, the average grade of classrooms with lowest grading standards is significantly higher than the one of classrooms with highest grading standards (9.1 against 7.8). These classrooms, with both lowest and highest GS, are the ones for which the distance between INVALSI score and grade is bigger: ideally, in a continuous that goes from the strictest teacher to the most generous teacher, it is possible to imagine a classroom for which the distance between INVALSI score and grade is null. The same identical patterns happen considering grading standards in Mathematics.

Table 2.1: Bottom/top 8 classrooms with teachers having lower/higher GS in Language and Mathematics, and respective classroom average of INVALSI test score and classroom average grade (N=9370).

	Language			***Mathematics***		
	Grading Standard (std)	INVALSI score (classroom average)	Grade (classroom average)	Grading Standard (std)	INVALSI score (classroom average)	Grade (classroom average)
Classrooms with lowest GS						
	-2.59	173.9	9.1	-2.44	162	9.1
	-2.36	182.7	9.4	-2.28	172	9.5
	-2.27	172.7	8.8	-2.2	163.3	9.4
	-2.16	177	8.9	-2.15	145.4	8.1
	-2.15	182.2	9.4	-2.11	167	8.9
	-1.98	195.1	9.6	-1.97	158.9	8.3
	-1.94	186	9.2	-1.95	156.7	8.3
	-1.90	165.3	8.4	-1.83	173.6	9
Mean	**-2.17**	**179.4**	**9.1**	**-2.12**	**162.4**	**8.8**
Classrooms with highest GS						
	2.06	233.5	7.1	2.56	293.1	8.9
	2.08	258.9	8.3	2.59	272.9	7.8
	2.26	243.1	7.3	2.61	281.5	8.1
	2.32	249.9	7.6	2.62	287.8	8.3
	2.39	264.3	7.9	2.72	271.1	7.5
	2.67	259	7.4	3.56	296.9	7.6
	2.89	269.6	8.1	3.94	313.2	7.4
	4.11	317.9	8.9	4.15	316.3	7.7
Mean	**2.60**	**262**	**7.8**	**3.2**	**291.6**	**7.9**

Note: INVALSI test score and grade are shown in their original scale: INVALSI score has mean 200 and S.D. 40; grades are in a scale from 1 to 10.

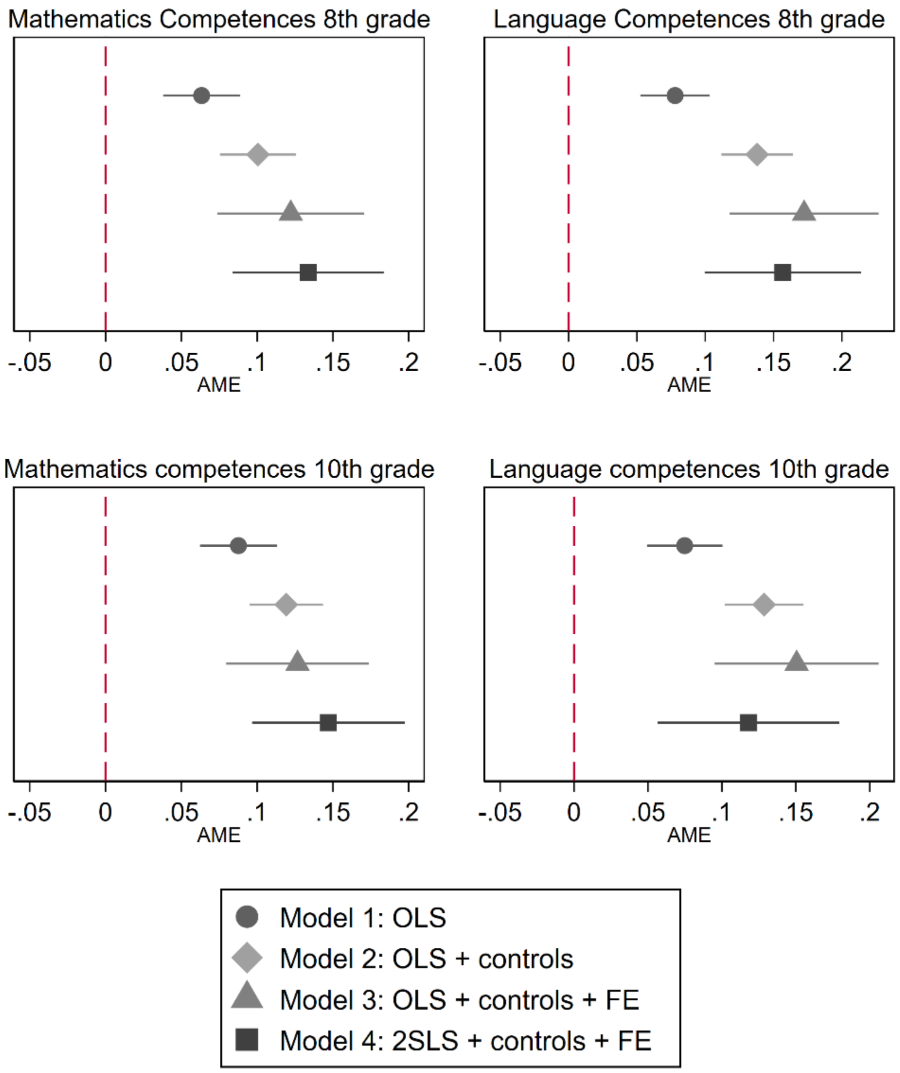
Effect of GS on Student Competences

In this section I report the findings related to the effect of grading standards on student competences in subsequent educational levels. Figure 2.2 reports the average marginal effects of teacher grading standards in 5th grade on INVALSI test score in 8th grade and in 10th grade in both Mathematics and Language, derived from four different models for each subject. The first model, which includes the treatment alone, shows that an increase of one standard deviation (SD, hereafter) in teacher grading standards corresponds to an increase of about 0.08 SDs in Language competences, both in 8th and 10th grade. For mathematics competences, one standard deviation in teacher grading standards is associated to a variation of 0.06 SDs in competences in grade 8 and nearly of 0.10 SDs in competences in grade 10.

Comparing the specification of model 1 to the specification of model 2, where students' demographic characteristics and previous ability are included, the coefficients increase for both subjects. In model 3 and 4, where fixed effects at the school level and the instrumental variable approach are adopted, we observe that an increase of one SD in teacher grading standards corresponds to an increase of about 0.15 SDs in students Language and Mathematics competences at the end of lower secondary education (grade 8), and in Mathematics competences in upper secondary education (grade 10). The increase in Language competences in grade 10 is slightly smaller, around 0.12 SDs.

In order to understand the magnitude of the increase in competences, results can be interpreted on the original scale of the INVALSI test score. The average result in INVALSI is around 200 points, with a standard deviation of 40.

Figure 2.2: Average marginal effects of GS in 5th grade on INVALSI test score in 8th and in 10th grade in Mathematics and Language competences; coefficients derived from OLS; N = 9370; 95% C.I.



Note: Model 1 controls for treatment. Model 2 includes students' sociodemographic and previous performance, teacher characteristics and classroom composition. Model 3 includes school fixed effect. Model 4 includes the instrumental variable.

Mathematics: F test instrument = 7898.05; Prob > F = 0.000

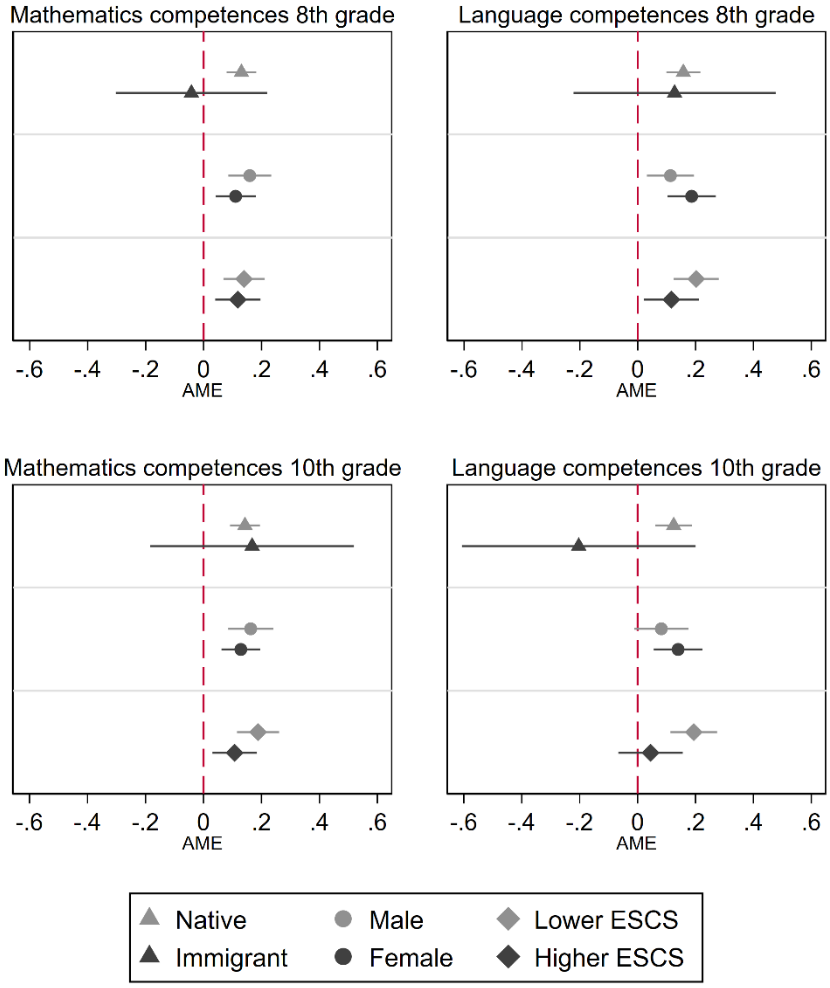
Language: F test instrument = 6698.02; Prob > F = 0.000

An increase of 1 standard deviation in grading standards correspond to an increase of about 6 points in the INVALSI test for both mathematics and language competences in 8th grade, and of about 5 to 6 points in 10th grade competences. All the model specifications suggest that 5th grade students who are exposed to a teacher with higher grading standards, or to a more severe teacher, are more likely to benefit in terms of competences gained three and five years later.

The next goal is understanding whether such positive impact of having a stricter teacher is similar or equal for students with different socio-demographic characteristics, therefore coming from different socioeconomic background, with opposite gender or with a migratory background or not. Importantly, since in this analysis we adjust for teachers' grades in the 4th grade, what we are looking at is the possible heterogeneous reactions to being exposed to certain grading standards across categories of students identified by ascriptive characteristics but with comparable levels of previous academic performance. Figure 2.3 shows the average marginal effects of grading standards on students' competences measured later on in time, by students' migratory background, gender and socioeconomic background. Coefficients are derived from model 4, with all control variables, fixed effects at the school level and the IV specification.

Results show that the positive effect of grading standards on students' Language and Mathematics competences is similar across students with different gender, migration background and social origin. The effect sizes are in most of the cases very similar and the 95% confidence intervals are widely overlapped. Concerning migratory background, instead, the inspection of effect sizes suggests a potential negative effect of high language grading standards for immigrant students in high school and a null effect for lower social background students. Unfortunately, the wide confidence intervals, make it difficult to provide a firm conclusion on these results based on our data.

Figure 2.3: Average marginal effects of GS in 5th grade on INVALSI test score in 8th and in 10th grade in Mathematics and Language competences by student characteristics: immigrant status, gender, ESCS; coefficients derived from OLS; N=9370; 95% C.I.



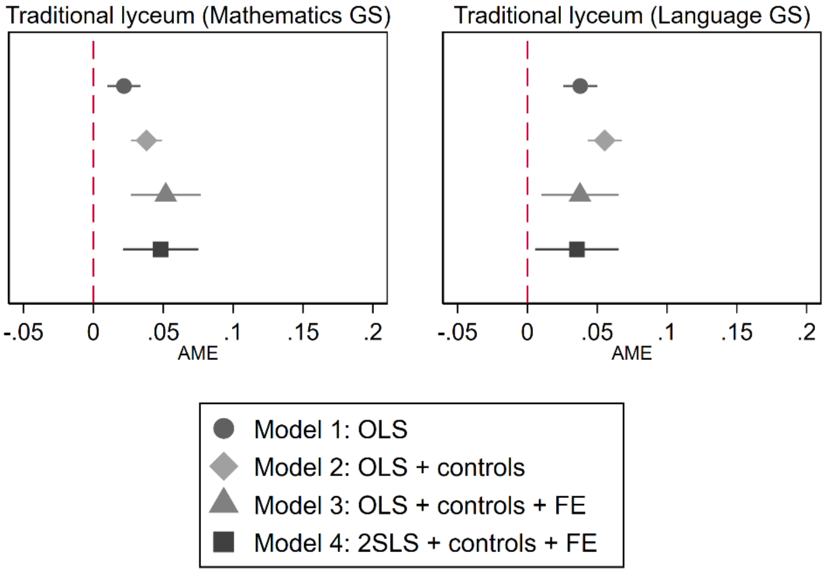
Note: Coefficients derived from model 4 (all control variables, fixed effect at the school level, iv specification)

Effect of GS on Student Probability of Being Enrolled in a Traditional Lyceum

In this section, the impact of teacher grading standards in 5th grade on students' probability of being enrolled in a traditional lyceum in 10th grade, rather than being enrolled in other lyceums, technical or vocational schools, is shown. In the analyzed sample, 38% of students are enrolled in traditional lyceums in grade 10. Figure 2.4 shows a positive effect of having a stricter teacher in 5th grade on the probability of being in a traditional lyceum rather than a non-traditional lyceum, or in a vocational or a technical school. In the baseline model specification without covariates (model 1), an increase of 1 standard deviation in teacher grading standards corresponds to an increase of 2 percentage points in the probability of being enrolled in lyceum having strict mathematics teacher, and of 4 percentage points in the probability of being enrolled in traditional lyceum having strict language teachers.

When including students' sociodemographic characteristics and previous ability, the effects slightly increase. There are no substantive differences between model 3 (with school fixed effects) and model 4 (iv specification). The effect of an increase of one standard deviation in the strictness of mathematics teacher in 5th grade corresponds to an increase of 5 percentage points in the probability of being enrolled in a traditional lyceum in 10th grade. Considering teacher grading strictness in language, the increase in the probability is 4 percentage points.

Figure 2.4: Average marginal effects of GS in 5th grade in Language and Mathematics on the probability of being enrolled in a traditional lyceum in 10th grade; coefficients derived from OLS; N = 9370; 95% C.I.

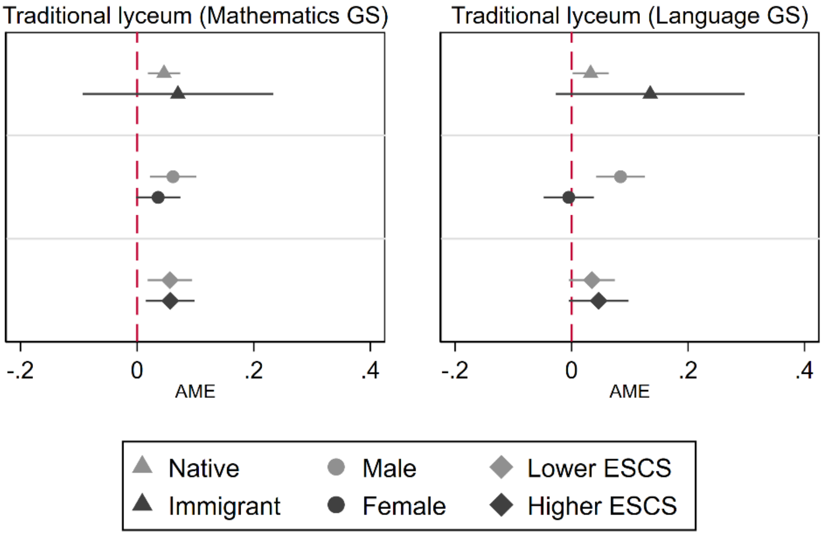


Note: Model 1 controls for treatment. Model 2 includes students' sociodemographic and previous performance, teacher characteristics and classroom composition. Model 3 includes school fixed effect. Model 4 includes the instrumental variable.
Mathematics: F test instrument = 7898.05; Prob > F = 0.000
Language: F test instrument = 6698.02; Prob > F = 0.000

The investigation of heterogeneous effects for students with different migratory background, gender and socioeconomic background is presented in Figure 2.5. Results show that having a mathematics teacher with higher grading standards at the end of primary education has a positive effect on the chances of being enrolled in a traditional lyceum 5 years later, and this effect is similar across students with different sociodemographic characteristics, but comparable early academic performance. The exception are immigrant students, for which the coefficient is not statistically significant probably because of the low sample size. Results are more controversial when considering language teachers.

It appears that immigrant students may benefit more from having a strict teacher in language in 5th grade comparing to native students, for which the effect is close to zero. Female students do not benefit in terms of enrollment in traditional lyceum from having had a strict teacher in Language in 5th grade compared to male students. Finally, looking at heterogeneous effect of grading standards, results indicate that students' ESCS does not moderate the positive effect of Language teacher grading standards on the probability of being enrolled in a traditional lyceum.

Figure 2.5: Average marginal effects of GS in 5th grade in Language and Mathematics on the probability of being enrolled in a traditional lyceum in 10th grade by student characteristics: immigrant status, gender, ESCS; coefficients derived from OLS; N=9370; 95% C.I.



Note: Coefficients derived from model 4 (all control variables, fixed effect at the school level, iv specification)

Conclusion and Discussion

This chapter addressed the issue of teacher grading standards in primary school, and how they affect important educational outcomes. The focus is on children's competences development and enrollment in academic tracks such as traditional lyceums in Italian schools. Grading standards is a measure reflecting of how strict the teacher is when evaluating and assigning grades to their students. Specifically, grading standards reflect the level of students' competences needed in order to get a specific grade, therefore students with similar competence but belonging to different classrooms may get higher grades when their teacher has lower grading standards and vice versa. Previous results suggest that through grading practices, and grading standards, teachers can manipulate students' effort and motivation: higher standards may induce students to increment their effort in order to satisfy teachers' requirements if they aspire to get a good grade, and, consequently, students can boost their competences development (Betts & Grogger 2003; Iacus & Porro 2008) and more generally they can benefit in terms of educational outcomes and choices. On the other hand, if teachers have grading standards that are too high to reach, it can induce students to give up, and this may have a detrimental effect on students' educational outcomes.

In line with most previous empirical findings (see Montmarquette and Mahseredjian 1989 for an exception), results show a positive effect of grading standards measured in primary school (5th grade) on both subject specific competences and probability of being enrolled in a traditional lyceum in high school. Results hold considering competences in Language and Mathematics and looking at different time points – three and five years after the treatment.

When looking at heterogeneous effect, results are less clear-cut. Concerning students' competences development throughout the years, it seems that 1st and 2nd generation immigrant students benefit less than native students from having a strict teacher, in both Language and Mathematics. Even if it is difficult to interpret results based on the estimates because of the low sample size of immigrant students, they may suggest that the effect for immigrant students is nearly zero, or even negative considering Language competences measured in 10th grade. This may be partially explained by the struggle that especially 1st generation immigrant students face in learning a new language, and having a strict teacher in primary school in Language may discourage them from learning and studying the subject, leading to detrimental consequences for their competences later in time. Concerning socioeconomic background, it seems that high ESCS students may benefit less from having a teacher with high grading standards in 5th grade, and the effect is null considering competences in Language measured in 10th grade. Focusing on the probability of being enrolled in a traditional lyceum, the positive effect of having a mathematics teacher with high grading standards in primary school is similar across students with different sociodemographic characteristics. Instead, the effect of having a strict language teacher is less straightforward: the effect is no longer significant looking at female students compared to male students, and looking only at students' socioeconomic background, but it becomes larger considering immigrant students compared to native students. Overall, despite such minor signs of heterogeneity based on migration background, our main conclusion is that in the Italian context, higher grading standards seem to have positive or at best null impacts on a variety of students' outcomes in lower and upper secondary education. I did not detect clear evidence for specific detrimental consequences for specific categories of students identified based on their socio-demographic characteristics.

In line with previous results on the topic conducted mostly in the United States, this work suggests that stricter grades might be overall beneficial for students' subsequent educational outcomes in Italy, even when measured in primary school. Interestingly, empirical investigations focusing on grading practices in primary education are scarce, even if it is considered a crucial moment in students' educational journey in terms of competences development (Facchinello 2020). Indeed, adopting hard grading standards on 10 years old pupils may have strong implications that deserve particular attention. For instance, higher grading standards within a classroom imply increased inequalities among students, that can push pupils to benefit from an early categorization and ranking of their abilities in comparison with their peers. This may have positive consequences on their motivation, self-esteem, self-identity, as well as on their endurance and effort, and consequently on their educational competences and trajectories. It is important to underline that this may hold in the analyzed context, in which relatively harder grading standards are in absolute terms not particularly hard, considering that grades of 5th grade pupils are overall high, and teachers are generally generous when attributing marks in this educational level.

This work presents some limitations. First, it is not possible to be completely sure about the validity of the selected instrument, even if it is exogenous to the explanatory variables, it correlates significantly to the explanatory variables, and the F tests hold the assumptions. However, the fact that results from different model specifications, including also the instrumental variable specifications, are pointing to similar estimates may suggest that the overall interpretation can hold. Second, The INVALSI test score may present some measurement error. Indeed, since it is a score measured only one time, while teacher grades are repeatedly assessed during the academic year, it may be biased due to different factors: the conditions in which the test was taken, students' specific emotional state or students' different proclivities associated with the belonging to specific groups regarding exams. However, it must be considered that INVALSI scores are weighted for different factors, and previous studies controlling for INVALSI measurement error (Lievore & Triventi 2022) suggest that results are not biased.

This work underlines how teachers should be aware of how specific grading practices, and particularly those considered as severe ones, may help their students, independently of students' sociodemographic characteristics. Following the work of Facchinello (2020), this chapter suggests that the social scientists should dedicate more attention on the topic of the grading system, particularly in the early stages of the educational career, in order to investigate aspects that have been somehow overlooked by the educational literature and might have important implications also in terms of educational policy making. This is especially true in the Italian context, in which the 2021 reform on primary schools eliminated numerical grades and promoted descriptive students' evaluations. Grading practices may have important effects on how students perceive themselves and their ability. Indeed, if students receive inflated grades – higher than what they deserve – their parents and themselves may believe that they are prepared for specific situations (e.g., highly demanding academic education), while they are not. Moreover, if very skilled and prepared students get the same grades as their less-prepared colleagues, this might instill a sense of frustration and demotivation in the former, thereby leading to reduced effort in schooling and participation in classroom activities (Finefter-Rosenbluh & Levinson 2015). In the long run, the entire work-ethic of students can result deteriorated from this process, since it may suggest that hard work is not needed for achieving educational success (Chowdhury 2018). This study shows that a reform of the grading practices in elementary school has been implemented without a careful consideration of the pros and cons and without a full consideration of the actual grading practices adopted by Italian teachers. These findings seem to suggest that in a context of overall generous evaluations towards children in primary education, adopting relatively stricter standards appear not to have negative consequences and, for most of students' categories, to positively affect their subsequent educational outcomes.

Appendix Chapter 2

Table A2.1: Description of the variables of interest

Control variables	**Coding and description**
Student sociodemographic & performance	
Gender	Recoded as 0 = Male; 1 = Female
Immigrant status	Recoded as 0 = Native; 1 = Immigrant I and II generation
ESCS	Standardized index from INVALSI composed by: parental occupation status, parental level of education, possession of specific material assets
Quarter of birth	Recoded as 0 = 1st quarter; 1 = 2nd quarter; 2 = 3rd quarter; 3 = 4th quarter
Regularity in studies	Recoded as 0 = Regular/early entrance; 1 = Late entrance
Attendance to infant school	Recoded as 0 = Yes; 1 = No; 2 = Missing
Attendance to kindergarten	Recoded as 0 = Yes; 1 = No; 2 = Missing
Student previous performance in (subject)	Grade at the end of 4th grade, self-reported (scale from 0 = 5 or less, to 5 = 10)
Teacher characteristics	
Gender	Recoded as 0 = Male; 1 = Female
Age	Continuous variable (scale from 25 to 68 in Mathematic; scale from 26 to 68 in Language)
Within-school seniority	Continuous variable (scale from 0 to 42 for Mathematics; scale from 0 to 41 for Language)
Educational credentials	Recoded as 0 = Teaching diploma; 1 = Bachelor/master degree/PhD
Parental education	Recoded as 0 = Lower; 1 = Higher
Type of contract	Recoded as 0 = Fixed-term; 1= Permanent
Teaching to test INVALSI (homework)	Recoded as 0 = No; 1 = Yes
Teaching to test INVALSI (in class)	Recoded as 0 = No; 1 = Yes
Classroom composition	
Share of female students	Continuous variable (scale from 0 to 100)
Mean ESCS (net of individual)	Standardized continuous variable
Classroom size	Continuous variable (scale from 11 to 29)
Share of immigrant students	Continuous variable (scale from 0 to 100)

Table A2.2: Balancing Tests: as-good-as-random distribution of Grading Standards across students in Language

	Gender		*Ethnic status*		*Socio-economic origin*	
Grading Standard in *Language*	-0.01	0	0	0	0.07**	0.07**
	(0.01)	(0.02)	(0.00)	(0.01)	(0.01)	(0.03)
Constant	0.52***	0.52***	0.06***	0.06***	0.16***	0.16***
	(0.01)	(0.01)	(0.00)	(0.00)	(0.01)	(0.01)
R-sqr	0	0.05	0	0.19	0	0.26
F-Statistic	0.29	0.89	0.64	0.63	0	0.01
BIC	13603.42	13104.42	251.51	-1711.51	25645.19	22844.7
AIC	13589.12	13090.13	237.22	-1725.8	25630.9	22830.41
Obs.	9370	9370	9370	9370	9370	9370
School FE	NO	YES	NO	YES	NO	YES

Table A2.3: Balancing Tests: as-good-as-random distribution of Grading Standards across students in Mathematics

	Gender		*Ethnic Status*		*Socio-economic origin*	
Grading Standard in *Mathematics*	-0.01	-0.01	0	0	0.04**	0.06**
	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)	(0.02)
Constant	0.52***	0.52***	0.06***	0.06***	0.16***	0.16***
	(0.01)	(0.01)	(0.00)	(0.00)	(0.01)	(0.01)
R-sqr	0	0.05	0	0.19	0	0.26
F-Statistic	0.08	0.42	0.73	0.58	0	0.01
BIC	13601.39	13103.76	251.61	-1711.58	25672.02	22844.67
AIC	13587.1	13089.47	237.32	-1725.87	25657.73	22830.38
Obs.	9370	9370	9370	9370	9370	9370
School FE	NO	YES	NO	YES	NO	YES

Table A2.4: Balancing Tests: as-good-as-random distribution of Language teachers' characteristics across classrooms in Language

	Lang. teachers' gender		Lang. teachers' age		Lang. teachers school seniotity		Lang. teacher education		Lang. teacher parental education		Lang. teacher type of contract		Lang. teaching to test (class)		Lang. teaching to test (homew)	
Class size	0.00	-0.00	0.16	0.22	0.07	0.21	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.01
	(0.00)	(0.00)	(0.08)	(0.14)	(0.09)	(0.17)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.01)
Constant	0.96***	1.05***	48.70***	47.45***	13.97***	11.13***	0.24**	0.17	0.36***	0.34	0.87***	0.86***	0.58***	0.53***	0.08	0.03
	(0.03)	(0.06)	(1.52)	(2.70)	(1.75)	(3.16)	(0.09)	(0.16)	(0.09)	(0.18)	(0.05)	(0.08)	(0.09)	(0.16)	(0.07)	(0.13)
R-sqr	0.00	0.49	0.01	0.60	0.00	0.58	0.00	0.55	0.00	0.55	0.00	0.59	0.00	0.61	0.00	0.58
F-Statistic	0.53	0.21	0.05	0.13	0.44	0.21	0.74	0.61	0.97	0.91	0.11	0.35	0.14	0.26	0.27	0.32
BIC	-643.12	-685.76	4950.57	2691.72	5156.51	2833.13	874.36	179.00	988.14	264.70	-43.48	-417.67	909.38	152.88	593.73	-9.24
Mean ESCS	-0.01	0.01	0.36	1.07	-0.25	0.72	0.06*	0.08	0.07*	0.04	-0.01	0.02	-0.03	-0.01	0.04	0.01
	(0.01)	(0.03)	(0.55)	(1.33)	(0.63)	(1.55)	(0.03)	(0.08)	(0.03)	(0.09)	(0.02)	(0.04)	(0.03)	(0.08)	(0.03)	(0.06)
Constant	0.98***	0.98***	51.55***	51.48***	15.30***	15.03***	0.27***	0.25***	0.36***	0.36***	0.94***	0.94***	0.71***	0.71***	0.16***	0.16***
	(0.01)	(0.01)	(0.29)	(0.33)	(0.34)	(0.39)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)	(0.02)
R-sqr	0.00	0.49	0.00	0.60	0.00	0.57	0.01	0.55	0.01	0.55	0.00	0.59	0.00	0.61	0.00	0.58
F-Statistic	0.49	0.64	0.52	0.42	0.69	0.64	0.04	0.31	0.03	0.67	0.67	0.70	0.38	0.94	0.15	0.93
BIC	-643.19	-682.97	4953.88	2694.99	5156.94	2835.83	870.28	177.42	983.61	264.35	-41.12	-416.18	910.80	155.39	592.85	-7.24
Perc immigr	-0.00*	0.00	-0.08***	-0.01	-0.04	0.02	0.00	-0.00	0.00	0.00	-0.00	0.00	0.00	0.01	0.00	-0.00
	(0.00)	(0.00)	(0.02)	(0.06)	(0.02)	(0.07)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.99***	0.97***	52.32***	51.61***	15.71***	14.86***	0.25***	0.27***	0.34***	0.34***	0.95***	0.93***	0.71***	0.65***	0.15***	0.17***
	(0.01)	(0.01)	(0.36)	(0.60)	(0.42)	(0.71)	(0.02)	(0.04)	(0.02)	(0.04)	(0.01)	(0.02)	(0.02)	(0.03)	(0.02)	(0.03)
R-sqr	0.01	0.49	0.02	0.59	0.00	0.57	0.00	0.55	0.00	0.55	0.01	0.59	0.00	0.62	0.00	0.58
F-Statistic	0.03	0.82	0.00	0.82	0.08	0.75	0.09	0.58	0.18	0.53	0.05	0.60	0.97	0.07	0.78	0.53
BIC	-647.55	-682.64	4941.53	2696.21	5154.06	2836.05	871.67	178.88	986.36	263.94	-44.67	-416.45	911.56	148.58	594.87	-8.02
Perc female	-0.00	-0.00	0.00	-0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	0.00	0.00
	(0.00)	(0.00)	(0.03)	(0.04)	(0.03)	(0.05)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.99***	1.06***	51.41***	52.71***	15.12***	14.88***	0.27***	0.23	0.33***	0.35**	0.92***	0.96***	0.78***	0.82***	0.06	0.04
	(0.03)	(0.05)	(1.33)	(2.07)	(1.54)	(2.42)	(0.08)	(0.12)	(0.08)	(0.14)	(0.04)	(0.06)	(0.08)	(0.12)	(0.06)	(0.10)
R-sqr	0.00	0.50	0.00	0.60	0.00	0.57	0.00	0.55	0.00	0.55	0.00	0.59	0.00	0.61	0.00	0.58
F-Statistic	0.70	0.06	0.90	0.56	0.91	0.94	0.98	0.87	0.75	0.97	0.65	0.71	0.35	0.33	0.11	0.22
BIC	-642.87	-689.79	4954.29	2695.61	5157.09	2836.25	874.47	179.46	988.04	264.72	-41.13	-416.17	910.70	153.46	592.35	-10.21
Obs.	712	446	712	446	712	446	712	446	712	446	712	446	712	446	712	446
School FE	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES

Table A2.5: Balancing Tests: as-good-as-random distribution of Mathematics teachers' characteristics across classrooms in Mathematics

	Math teachers' gender		**Math teachers' age**		**Math teacher school seniority**		**Math teacher education**		**Math teacher parental education**		**Math teacher type of contract**		**math teaching to test (class)**		**math teaching to test (homew)**	
Class size	0.00	0.00	0.09	0.18	0.02	0.13	0.00	-0.01	0.00	-0.01	0.00	0.01	0.00	0.01	0.00	0.01
	(0.00)	(0.00)	(0.09)	(0.16)	(0.09)	(0.16)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.01)
Constant	0.96***	0.96***	49.61***	47.77***	14.26***	12.17***	0.25**	0.43**	0.34***	0.51**	0.89***	0.84***	0.69***	0.61***	0.10	-0.03
	(0.03)	(0.06)	(1.63)	(2.89)	(1.73)	(2.99)	(0.09)	(0.16)	(0.09)	(0.18)	(0.05)	(0.09)	(0.09)	(0.16)	(0.07)	(0.14)
R-sqr	0.00	0.49	0.00	0.57	0.00	0.62	0.00	0.61	0.00	0.56	0.00	0.55	0.00	0.58	0.00	0.55
F-Statistic	0.80	0.71	0.30	0.24	0.85	0.40	0.74	0.32	0.83	0.41	0.27	0.24	0.88	0.53	0.37	0.14
BIC	-496.45	-772.30	5051.85	2752.51	5136.39	2784.12	888.14	147.17	990.51	263.34	19.88	-376.18	923.26	178.71	628.95	62.25
Mean ECSC	0.01	0.05	0.09	1.15	-1.79**	1.44	0.09**	-0.08	0.03	-0.01	-0.01	0.01	-0.04	-0.04	0.04	0.02
	(0.01)	(0.03)	(0.59)	(1.42)	(0.62)	(1.47)	(0.03)	(0.08)	(0.03)	(0.09)	(0.02)	(0.04)	(0.03)	(0.08)	(0.03)	(0.07)
Constant	0.97***	0.98***	51.27***	51.13***	14.67***	14.63***	0.27***	0.28***	0.36***	0.36***	0.94***	0.94***	0.70***	0.71***	0.17***	0.18***
	(0.01)	(0.01)	(0.31)	(0.35)	(0.33)	(0.36)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)	(0.02)
R-sqr	0.00	0.50	0.00	0.57	0.01	0.62	0.01	0.61	0.00	0.55	0.00	0.55	0.00	0.58	0.00	0.55
F-Statistic	0.50	0.08	0.88	0.42	0.00	0.33	0.00	0.29	0.39	0.94	0.41	0.83	0.23	0.57	0.11	0.78
BIC	-496.83	-778.21	5052.92	2753.99	5128.15	2783.59	879.79	146.95	989.80	264.72	20.41	-373.48	921.85	178.88	627.23	66.44
Perc immigr	-0.00	-0.00	-0.09***	-0.03	-0.03	0.02	0.00**	0.00	0.00	0.00	-0.00***	0.00	-0.00	0.01**	0.00	0.00
	(0.00)	(0.00)	(0.02)	(0.06)	(0.02)	(0.06)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.97***	0.99***	52.12***	51.39***	14.92***	14.48***	0.24***	0.26***	0.35***	0.36***	0.97***	0.94***	0.71***	0.63***	0.16***	0.16***
	(0.01)	(0.01)	(0.38)	(0.64)	(0.41)	(0.67)	(0.02)	(0.03)	(0.02)	(0.04)	(0.01)	(0.02)	(0.02)	(0.04)	(0.02)	(0.03)
R-sqr	0.00	0.49	0.02	0.57	0.00	0.62	0.01	0.61	0.00	0.55	0.03	0.55	0.00	0.60	0.00	0.55
F-Statistic	0.62	0.16	0.00	0.67	0.18	0.75	0.01	0.43	0.29	0.94	0.00	0.98	0.65	0.01	0.58	0.64
BIC	-496.63	-776.04	5039.10	2754.95	5134.60	2785.32	881.42	147.90	989.43	264.72	-0.61	-373.39	923.08	164.82	629.44	66.16
Perc female	0.00	-0.00*	-0.01	0.02	0.02	0.02	0.00	0.00	0.00	-0.00	0.00	-0.00	-0.00	-0.01*	0.00	0.00
	(0.00)	(0.00)	(0.03)	(0.04)	(0.03)	(0.05)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.97***	1.07***	51.95***	50.41***	13.77***	13.70***	0.19*	0.21	0.28***	0.37**	0.93***	1.01***	0.82***	0.98***	0.16*	0.14
	(0.03)	(0.04)	(1.43)	(2.21)	(1.51)	(2.29)	(0.08)	(0.12)	(0.08)	(0.14)	(0.04)	(0.07)	(0.08)	(0.12)	(0.06)	(0.11)
R-sqr	0.00	0.50	0.00	0.57	0.00	0.62	0.00	0.61	0.00	0.55	0.00	0.55	0.00	0.59	0.00	0.55
F-Statistic	0.99	0.04	0.63	0.73	0.58	0.67	0.24	0.55	0.31	0.95	0.86	0.32	0.11	0.03	0.92	0.76
BIC	-496.38	-780.86	5052.70	2755.08	5136.12	2785.17	886.84	148.45	989.52	264.72	21.06	-375.37	920.75	169.77	629.75	66.40
Obs.	712	446	712	446	712	446	712	446	712	446	712	446	712	446	712	446
School FE	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES

Table A2.6: Comparison between linear regression models and quadratic regression models predicting INVALSI test score in Language and Mathematics in 8th grade. Coefficients derived from model 3 (all controls + fixed effects at the school level). Standard error in parentheses; ***p<0.01, **p<0.05, *p<0.1

	Language		Language		Mathematics		Mathematics	
	8th grade	(S.E.)	8th grade	(S.E.)	8th grade	(S.E.)	8th grade	(S.E.)
Grading Standards (5th grade)	0.133***	(0.029)	0.145***	(0.030)	0.090***	(0.027)	0.108***	(0.029)
Grading Standards ^2			-0.026	(0.016)			**-0.028***	(0.016)
Student Characteristics								
Female (Ref. Male)	0.263***	(0.020)	0.263***	(0.020)	-0.183***	(0.021)	-0.181***	(0.021)
Quarter of birth (Ref. 1st)								
2nd quarter	0.073***	(0.028)	0.074***	(0.028)	0.067**	(0.028)	0.067**	(0.028)
3rd quarter	-0.006	(0.028)	-0.005	(0.028)	-0.018	(0.028)	-0.017	(0.028)
4th quarter	0.004	(0.028)	0.004	(0.028)	0.000	(0.029)	-0.000	(0.029)
ESCS	0.215***	(0.012)	0.216***	(0.012)	0.168***	(0.012)	0.167***	(0.012)
Immigrant (Ref. Native)	-0.182***	(0.045)	-0.180***	(0.045)	-0.106**	(0.046)	-0.107**	(0.046)
Late entrance (Ref. Regular)	-0.122	(0.091)	-0.122	(0.091)	-0.060	(0.093)	-0.060	(0.093)
Attendance to infant school (Ref. Yes)								
No	-0.005	(0.025)	-0.005	(0.025)	0.003	(0.026)	0.003	(0.026)
Missing	-0.028	(0.050)	-0.027	(0.050)	-0.044	(0.051)	-0.044	(0.051)
Attendance to kindergarten (Ref. Yes)								
No	-0.123*	(0.067)	-0.120*	(0.067)	-0.170**	(0.068)	-0.170**	(0.068)
Missing	-0.738***	(0.209)	-0.737***	(0.209)	-0.597***	(0.212)	-0.600***	(0.212)
Teacher Characteristics								
Female (Ref. Male)	-0.134	(0.118)	-0.136	(0.118)	-0.300**	(0.130)	-0.260**	(0.132)
Age	0.001	(0.003)	0.002	(0.003)	0.002	(0.003)	0.003	(0.003)
Seniority in school (years)	-0.005**	(0.003)	-0.006**	(0.003)	-0.004	(0.003)	-0.004	(0.003)
Bachelor/Master/PhD (Ref. Teaching diploma)	0.003	(0.044)	0.005	(0.044)	-0.122***	(0.044)	-0.127***	(0.044)
Parental education higher (Ref. Lower)	-0.005	(0.039)	-0.005	(0.039)	0.012	(0.038)	0.013	(0.038)
Permanent contract (Ref. Fixed-term)	0.311***	(0.092)	0.309***	(0.092)	0.033	(0.091)	0.038	(0.091)
Teaching to test in class yes (Ref. No)	-0.032	(0.045)	-0.031	(0.045)	-0.058	(0.044)	-0.056	(0.044)
Teaching to test homework yes (Ref. No)	-0.004	(0.053)	-0.010	(0.053)	0.028	(0.049)	0.024	(0.049)
Classroom Composition								
% Female	-0.001	(0.001)	-0.001	(0.001)	0.001	(0.001)	0.002	(0.002)
Mean ESCS	0.023	(0.053)	0.027	(0.053)	-0.045	(0.054)	-0.057	(0.054)
Class size	-0.002	(0.006)	-0.002	(0.006)	-0.002	(0.006)	-0.002	(0.006)
% Immigrants	0.001	(0.002)	0.001	(0.002)	-0.001	(0.002)	-0.002	(0.002)
Constant	-0.074	(0.247)	-0.088	(0.247)	0.456*	(0.242)	0.413*	(0.243)
Observations	9,370		9,370		9,370		9,370	
R-squared	0.191		0.191		0.202		0.202	
AIC	24351.71		24350.88		24658.88		24657.37	
BIC	24530.34		24536.65		24837.51		24843.14	

Table A2.7: Comparison between linear regression models and quadratic regression models predicting INVALSI test score in Language and Mathematics in 10th grade. Coefficients derived from model 3 (all controls + fixed effects at the school level). Standard error in parentheses; ***p<0.01, **p<0.05, *p<0.1

	Language		Language		Mathematics		Mathematics	
	10th grade	(S.E.)	10th grade	(S.E.)	10th grade	(S.E.)	10th grade	(S.E.)
Grading Standards (5th grade)	0.114***	(0.029)	0.136***	(0.030)	0.097***	(0.026)	0.105***	(0.028)
Grading Standards ^2			**-0.052*****	(0.016)			-0.013	(0.015)
Student Characteristics								
Female (Ref. Male)	0.251***	(0.020)	0.251***	(0.020)	-0.219***	(0.020)	-0.218***	(0.020)
Quarter of birth (Ref. 1st)								
2nd quarter	0.045*	(0.027)	0.047*	(0.027)	0.021	(0.027)	0.021	(0.027)
3rd quarter	-0.033	(0.027)	-0.032	(0.027)	-0.039	(0.027)	-0.039	(0.027)
4th quarter	0.020	(0.028)	0.021	(0.028)	-0.021	(0.028)	-0.021	(0.028)
ESCS	0.244***	(0.012)	0.245***	(0.012)	0.216***	(0.011)	0.216***	(0.011)
Immigrant (Ref. Native)	-0.142***	(0.045)	-0.137***	(0.045)	-0.077*	(0.044)	-0.078*	(0.044)
Late entrance (Ref. Regular)	-0.133	(0.091)	-0.133	(0.091)	-0.050	(0.090)	-0.049	(0.090)
Attendance to infant school (Ref. Yes)								
No	-0.009	(0.025)	-0.009	(0.025)	-0.010	(0.025)	-0.010	(0.025)
Missing	0.069	(0.050)	0.071	(0.050)	0.051	(0.049)	0.051	(0.049)
Attendance to kindergarten (Ref. Yes)								
No	-0.122*	(0.066)	-0.116*	(0.066)	-0.154**	(0.066)	-0.154**	(0.066)
Missing	-0.831***	(0.207)	-0.830***	(0.207)	-0.921***	(0.205)	-0.922***	(0.205)
Teacher Characteristics								
Female (Ref. Male)	-0.162	(0.117)	-0.167	(0.117)	-0.047	(0.125)	-0.028	(0.127)
Age	-0.002	(0.003)	0.000	(0.003)	-0.000	(0.003)	-0.000	(0.003)
Seniority in school (years)	0.000	(0.003)	0.000	(0.003)	0.000	(0.003)	-0.000	(0.003)
Bachelor/Master/PhD (Ref. Teaching diploma)	-0.037	(0.044)	-0.033	(0.044)	-0.086**	(0.043)	-0.088**	(0.043)
Parental education higher (Ref. Lower)	-0.006	(0.038)	-0.006	(0.038)	0.028	(0.037)	0.029	(0.037)
Permanent contract (Ref. Fixed-term)	0.043	(0.091)	0.040	(0.091)	0.014	(0.087)	0.017	(0.087)
Teaching to test in class yes (Ref. No)	0.038	(0.044)	0.039	(0.044)	-0.026	(0.043)	-0.025	(0.043)
Teaching to test homework yes (Ref. No)	-0.020	(0.052)	-0.031	(0.052)	-0.007	(0.048)	-0.009	(0.048)
Classroom Composition								
% Female	0.001	(0.001)	0.001	(0.001)	-0.002	(0.233)	-0.022	(0.235)
Mean ESCS	0.042	(0.052)	0.050	(0.052)	-0.002	(0.052)	-0.008	(0.052)
Class size	0.011**	(0.005)	0.010*	(0.005)	0.009*	(0.005)	0.009*	(0.005)
% Immigrants	-0.003	(0.002)	-0.002	(0.002)	-0.001	(0.002)	-0.001	(0.002)
Constant	-0.182	(0.245)	-0.209	(0.245)	-0.002	(0.233)	-0.022	(0.235)
Observations	9,370		9,370		9,370		9,370	
R-squared	0.197		0.198		0.231		0.231	
AIC	24172.41		24163.16		23971		23972.17	
BIC	24351.04		24348.93		24149.63		24157.95	

Table A2.8: Comparison between linear regression models and quadratic regression models predicting students' enrollment in traditional lyceums in 10th grade for Language grading standards and Mathematics grading standards. Coefficients derived from model 3 (all controls + fixed effects at the school level). Standard error in parentheses; ***p<0.01, **p<0.05, *p<0.1

	Trad. Lyceum (Lang)	(S.E.)	Trad. Lyceum (Lang.)	(S.E.)	Trad. Lyceum (Maths)	(S.E.)	Trad. Lyceum (Maths)	(S.E.)
Grading standards (5th grade)	0.027*	(0.014)	0.031**	(0.015)	0.041***	(0.013)	0.049***	(0.014)
Grading standards ^2			-0.009	(0.008)			-0.011	(0.007)
Student Characteristics								
Female (Ref. Male)	-0.026***	(0.010)	-0.026***	(0.010)	-0.026***	(0.010)	-0.025**	(0.010)
Quarter of birth (Ref. 1st)								
2nd quarter	0.024*	(0.013)	0.025*	(0.013)	0.024*	(0.013)	0.024*	(0.013)
3rd quarter	-0.000	(0.013)	-0.000	(0.013)	-0.001	(0.013)	-0.001	(0.013)
4th quarter	-0.007	(0.014)	-0.007	(0.014)	-0.007	(0.014)	-0.008	(0.014)
ESCS	0.138***	(0.006)	0.139***	(0.006)	0.139***	(0.006)	0.139***	(0.006)
Immigrant (Ref. Native)	-0.054**	(0.022)	-0.053**	(0.022)	-0.051**	(0.022)	-0.052**	(0.022)
Late entrance (Ref. Regular)	-0.044	(0.045)	-0.044	(0.045)	-0.045	(0.045)	-0.045	(0.045)
Attendance to infant school (Ref. Yes)								
No	-0.018	(0.012)	-0.018	(0.012)	-0.017	(0.012)	-0.017	(0.012)
Missing	-0.001	(0.025)	-0.001	(0.025)	-0.001	(0.025)	-0.001	(0.025)
Attendance to kindergarten (Ref. Yes)								
No	0.045	(0.033)	0.046	(0.033)	0.039	(0.033)	0.040	(0.033)
Missing	-0.197*	(0.102)	-0.197*	(0.102)	-0.197*	(0.102)	-0.198*	(0.102)
Teacher Characteristics								
Female (Ref. Male)	-0.135**	(0.058)	-0.136**	(0.058)	-0.058	(0.062)	-0.042	(0.063)
Age	0.000	(0.001)	0.001	(0.001)	0.002	(0.001)	0.002	(0.001)
Seniority in school (years)	0.000	(0.001)	0.000	(0.001)	-0.002	(0.001)	-0.002*	(0.001)
Bachelor/Master/PhD (Ref. Teaching diploma)	0.003	(0.021)	0.004	(0.021)	0.030	(0.021)	0.028	(0.021)
Parental education higher (Ref. Lower)	0.016	(0.019)	0.016	(0.019)	0.048***	(0.018)	0.048***	(0.018)
Permanent contract (Ref. Fixed-term)	0.068	(0.045)	0.067	(0.045)	-0.013	(0.043)	-0.011	(0.043)
Teaching to test in class yes (Ref. No)	-0.010	(0.022)	-0.010	(0.022)	-0.033	(0.021)	-0.032	(0.021)
Teaching to test homework yes (Ref. No)	-0.040	(0.026)	-0.042	(0.026)	-0.007	(0.024)	-0.009	(0.024)
Classroom Composition								
% Female	-0.000	(0.001)	-0.000	(0.001)	-0.000	(0.001)	-0.000	(0.001)
Mean ESCS	0.091***	(0.026)	0.092***	(0.026)	0.093***	(0.026)	0.088***	(0.026)
Class size	0.004	(0.003)	0.004	(0.003)	0.005*	(0.003)	0.005*	(0.003)
% Immigrants	0.001	(0.001)	0.001	(0.001)	0.001	(0.001)	0.001	(0.001)
Constant	0.360***	(0.120)	0.355***	(0.121)	0.273**	(0.116)	0.256**	(0.117)
Observations	9,370		9,370		9,370		9,370	
R-squared	0.211		0.212		0.212		0.212	
AIC	10875.12		10875.6		10866.46		10866.01	
BIC	11053.75		11061.38		11045.1		11051.79	

Figure A1.1: Predicted values derived from quadratic regression models predicting INVALSI test score in Mathematics in 8th grade. Coefficients derived from model 3 (all controls + fixed effects at the school level).

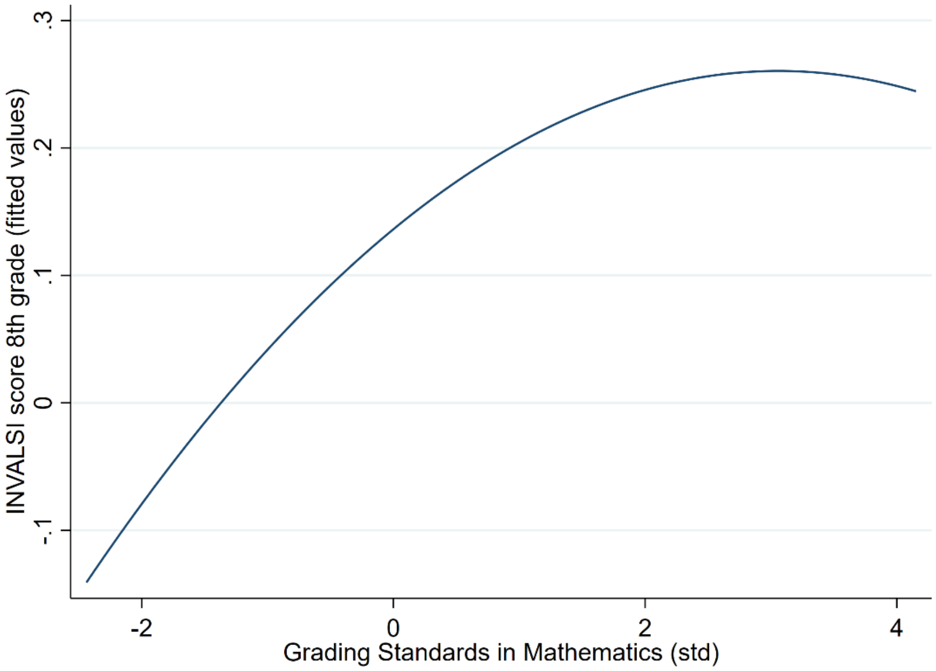
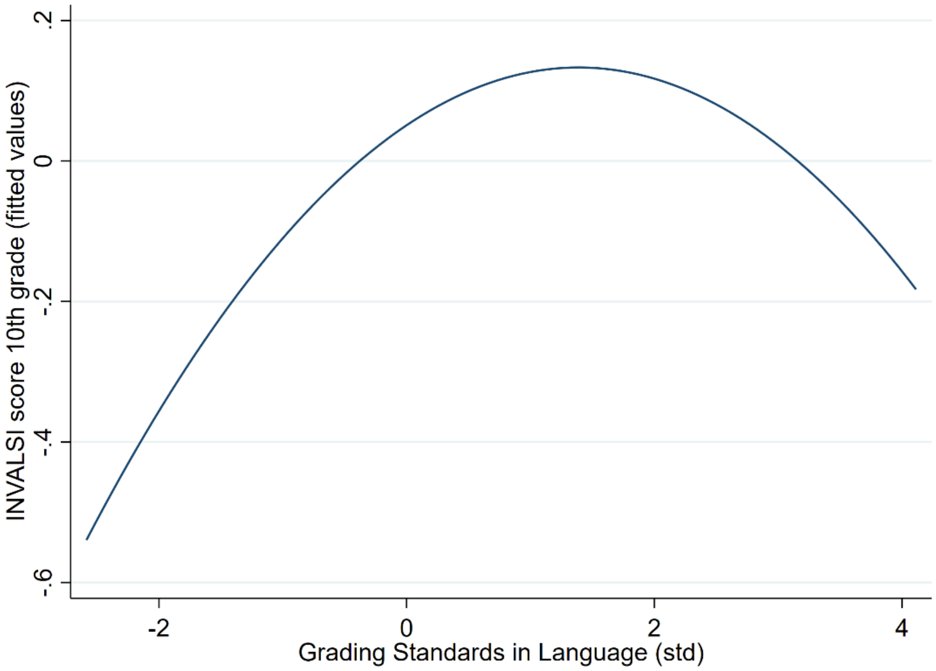


Figure A1.2: Predicted values derived from quadratic regression models predicting INVALSI test score in Language in 10th grade. Coefficients derived from model 3 (all controls + fixed effects at the school level).



Chapter 3

Do teacher and classroom characteristics affect the way in which girls and boys are graded? A multilevel analysis of student-teacher matched data[11]

Abstract

Teachers' evaluations of students do not consider only academic competence, but are imbued with social considerations related to individual teacher and student characteristics, their interactions, and the surrounding context. The aim of this chapter is understanding the extent to which teachers grade girls more generously than boys, and which characteristics of teachers and classrooms are likely to reduce this gender grading gap. I use Italian data from INVALSI-SNV, providing information on 10th-grade students linked with their teachers. The analysis relies on grade equation models in multilevel regression analysis, with students as first level, teachers/classrooms as second level, and schools as third level. Results show that, when comparing students who have identical subject-specific competence, teachers are more likely to give higher grades to girls. Furthermore, they demonstrate for the first time that this grading premium favouring girls is systemic, as teacher and classroom characteristics play a negligible role in reducing it.

Keywords: teachers' grades; gender; grading mismatch; academic performance; education; social inequalities

[11] Chapter 3 is co-authored with Moris Triventi and published in the *British Journal of Sociology of Education* (DOI: 10.1080/01425692.2022.2122942)

Introduction

During the last century, the existence of gender-related gaps in educational outcomes has been widely demonstrated almost worldwide (Goldin, Katz and Kuziemko 2006), and gender gaps in educational achievement are among most discussed topics in the field of educational studies. Typically, when comparing standardized test scores in most OECD countries, girls outperform boys in humanities, language, and reading skills, while boys show better scores in mathematics (PISA 2019; OECD 2019; IES 2009). Gender gaps in standardized test scores are even broader among students who demonstrate higher performance (OECD 2014).

At the same time, previous literature shows that, when skills are measured through grades, female students display higher educational attainment rates than male students in recent cohorts across all subjects (Downey and Yuan 2005). Female students outperform boys in terms of average school achievement, university enrolment (Becker, Hubbard, and Murphy 2010; Pekkarinen 2012), and tertiary degree attainment (Schofer and Meyer 2005). Therefore, even if men are still considered to be higher performing in scientific fields, and consequently women are underrepresented in highly financially rewarding professions (Card and Payne 2017), on average, girls have higher educational attainments. These findings have fostered the development of additional studies seeking to explain the sources of the disadvantages faced by boys in education, in particular by examining the roles of both the family and the school system (Mickelson 1989; Dumais 2002; Buchmann and DiPrete 2006).

Our work follows this stream of the literature, by aiming to improve the understanding of what role the school system, particularly teachers, plays in the production of gender differences in educational attainment. Some studies have shown that teachers assess girls' performances higher than boys' (Emanuelsson and Fischbein 1986); in addition, girls get significantly higher grades in both mathematics (Falch and Naper 2013) and language (Machin and Pekkarinen 2008; Falch and Naper 2013). However, from these studies it is not clear whether girls' higher grades stem from higher academic efforts and competence, or whether their marks reflect preferential treatment from their teachers.

This article investigates teacher grading – the way in which teachers assign marks to their students – as a potential source generating gender inequalities in education. Specifically, it examines whether teachers grade female students more generously compared to male students who have the same subject-specific competence, as measured via standardized test scores, and examines whether this putative gender grading premium varies according to key teacher characteristics, features of the classroom, and elements of the school environment.

Investigating the way in which teachers assess their students is important for several reasons. First, grades might affect students' motivation and effort in education, and therefore affect their subsequent educational outcomes. Moreover, grades are an indicator of students' academic capacity which the school sends to parents, which thus might affect familial decisions to invest more or less in a child's educational career. Teacher assessment is used and processed as an important source of information, not only by students and families, but also by teachers themselves, who base their educational recommendations on the prior indicators of student ability to which they have access: student grades. Grades are often determinant factors for students, in accessing the next level of education, finding a place among the best educational options, attaining rankings in the classroom, gaining a particular scholarship, or conquering admission to college

(Blossfeld et al. 2016; Bonesrønning 2004; Bonner and Chen 2019). Moreover, teachers' assessments might have also long-term consequences, influencing occupational choices and thereby earnings, in students' adulthood (Lavy and Sand 2015; Borghans et al. 2016; Bonner and Chen 2019).

During the last few years, some researchers have attempted to assess whether grading involves systemic biases related to students' gender, and recent studies have explored systemic differences among teachers in upper-secondary education female and male students who have the same levels of competence. In Israel, Lavy (2008) found that girls obtained higher grades than boys in 'non-blind' tests in which the evaluator did know the student's identity, compared to 'blind' tests. Likewise, Lindahl (2007) observed that, when comparing students with the same level of competence, teachers in Sweden assessed girls more generously than boys; Angelo (2014) also reported similar findings for Portuguese high schools, as did Terrier (2015) for French high schools. Enzi (2015), analysing upper secondary education in Germany, found that gender plays a role in grading, a result that echoes that of Kiss (2013) in studying lower educational levels. Furthermore, among 15-year-old Czech students, there is a sizeable gender gap in teachers' grading which favours female students, in both language and mathematics (Protivínský and Münich 2018). Although most studies have detected a grading premium favouring girl, one study of the Nordic countries did not find any gender gap in grading (Hinnerich, Höglin and Johannesson 2011).

Nonetheless, although this research stream provides important information on gender differences in education, it does not clarify whether biased grading is more likely to take place in some specific learning environments in comparison with others. This is particularly relevant from a sociological perspective (Reimer 2019), as teachers' grading behaviours not only take into account student competencies, but also reflect the social and the cultural contexts, the school environment (Farkas et al. 1990), the teachers' own beliefs (Chen and Bonner 2017), the relationship between students and their teacher (Costrell 1994), and teachers' sociodemographic characteristics and teaching approach (Bonner and Chen 2019).

This article aims to gain a more accurate understanding of gender-based differences in teachers' grading by focusing on an understudied context: Italian upper-secondary schools, in which gender differences in academic performance, as measured via teachers' evaluations, are particularly pronounced. Moreover, in Italy the teaching profession is granted by the Constitutional right of 'autonomy' in duty delivery, being characterized by a certain degree of autonomy from other teachers or even the school administration regarding educational choices as grading (Bracci 2009). Additionally, the Italian educational system is characterized by vast territorial divides in school resources and socio-cultural environments (Checchi 2004; Montanaro 2008). For these reasons, Italy can be considered a 'best case' scenario, in which heterogeneity in grading favouritism related to student gender is likely to occur.

The empirical part of the current research relies on novel student–teacher matched data, which has only recently become available in Italy and permits in-depth investigation regarding the issue of our interest. I use a sample of about 39,000 students from two cohorts who were enrolled in 10th grade in the academic years 2015–2016 and 2016–2017. By relying on hierarchical linear regression models, I assess not only the overall grading gap between female and male students in two key subjects (language and mathematics), but also the extent to which this gap varies according to teacher characteristics, classroom composition, and school tracks.

The chapter is organised as follows. The next section develops the theoretical framework and formulates potential explanations for the gender grading mismatch, and the third section provides relevant information about the Italian educational system. The fourth section includes information about the dataset, the variables, and the methodology. The fifth section presents the results of the empirical analysis, and finally, the last section discusses our findings and the conclusion.

The Sources of the Gender Grading Gap (GGG)

In this section, I develop a theoretical framework that may aid understanding of the mechanisms behind the gender grading gap (GGG). Teachers' grades result from a complex assessment process that incorporates multiple indicators of students' performance, reflecting not only students' levels of competence, effort and motivation (OECD 2013), but also other important aspects, such as teachers' own expectations, considerations and beliefs (Chen and Bonner 2017), and the setting in which the relationships between students and their teacher develops (Wright, Horn and Sanders 1997).

In order to understand why teachers may be likely to give female students higher grades, teacher expectations theory is a useful tool. Accordingly, teachers have specific perceptions of, and therefore expectations about, their students. Such expectations involve an *a priori* evaluation of student competencies in specific domains (Cooper and Tom 1984). Several studies have attempted to assess the associational relationship between specific students' characteristics and teacher perception or expectation bias (as recent examples, see Hornstra et al. 2018; Zhu, Urhahne and Rubie-Davies 2018; and

Riegle-Crumb and Humphries 2012). According to this prior literature, teacher expectations are strictly linked to stereotypes. Stereotypes, as representations of characteristics of specific groups (Bordalo et al. 2016), are means teachers use in order to process information about students in an easy and efficient manner. Together with students' ethnic background (see Strand 2012) and socioeconomic background (see Speybroeck et al. 2012), gender is one of the main sources generating teacher stereotypes. Regarding students' gender, internalised representations may result in biased judgements and in over or under-evaluation of specific students' performance. Teachers tend to perceive girls as more motivated, as more eager to learn (e.g., Anders, McElvany and Baumert 2010; Gentrup and Rjosk 2018), as more well-behaved (Glock and Kleen 2017), and as enacting less disruptive behaviour in the classroom (Terrier 2015). However, teachers may also have biased perceptions of gender-based relative talents for specific subjects, because, as research indicates, they are likely to believe maths is more difficult for girls than for identically performing boys (Riegle-Crumb and Humphries 2012). Therefore, teacher evaluation bias can result in a 'premium' or 'penalty' connected to dominant gender stereotypes. Thus, following previous research, I formulate the first hypothesis:

H1: Teachers are more likely to grade female students more generously than male students who have the same standardized test scores.

However, students' signals of their own cognitive abilities may be interpreted differently by teachers who have specific characteristics, or are operating in a given context. Some studies demonstrate how students benefit from having a same-gender teacher (Ammermueller and Dolton 2006). Accordingly, the 'stereotype threat' theory (Steel 1997) explains how the similarity between the demographic characteristics of students (such as gender) and those of their teachers improves communications and mutual understandings between teacher and student. This could lead teachers to unconsciously reward ways of behaving that are similar to their own. In this respect, it has been suggested that the increase in the share of female teachers may explain the gender gap in achievement that favours females, even if there is contrasting evidence on this topic (Neugebauer, Helbig and Landmann 2011).

Other teacher characteristics may influence the ways in which they grade their students according to gender. Carlana (2019) showed how the gender-science implicit association test – which measures gender stereotypes among teachers – correlates with some teachers' observable traits, such as their gender, field of study, or being the parents of daughters. It is reasonable to assume that other teacher characteristics may affect how they evaluate students, such as via age or seniority, as more experienced teachers could be less likely to be driven by stereotypes or expectations when evaluating students. Moreover, in general, teachers' working conditions may also influence their approach to grading students. For example, teachers' salaries, their level of stability in role, or their work opportunities (Basilio and Almeida 2018) may impact their investment in making high-quality assessments of student abilities, through a 'disgruntled worker' effect (Vegas and de Laat 2003). Therefore, I expect that overall teacher characteristics contribute to moderate the GGG. More specifically, I develop two hypotheses:

H2a, Resemblance hypothesis: Male teachers are less likely to over-grade girls compared to female teachers.

H2b, Experience hypothesis: Older teachers, teachers with more years of in-school seniority, and those with a permanent contract display a lower GGG.

Very few studies have investigated how classroom or school features might affect the differences in teachers' grading of boys versus that of girls. However, insights from previous educational studies suggest possible moderation effects. For example, it has been demonstrated that in unfavourable contexts, such as classrooms that contain a large number of overall students; a high percentage of minority students; or a high proportion of students from a low socio-economic background, disciplinary problems are more frequently present (Rindermann 2007). In these problematic environments, teachers are more likely to be involved in conflict resolution, and less likely to interact with their students. This may reduce the time invested in assessing student achievement (Hochweber, Hosenfeld and Klieme 2013), and may make teachers more prone to rely on their expectations.

This reasoning may also apply to the type of school, because vocational and technical schools are more likely to be attended by students with lower school outcomes and poor educational expectations (Panichella and Triventi 2014). In such contexts, the gender gaps in grading might be larger, as teachers have fewer resources to accurately evaluate their students.

Moreover, the share of female students in the classroom may matter. Indeed, on one hand, boys may benefit from being in a classroom including a large share of girls, simply because the learning environment in such classrooms may be easier to manage for teachers. On the other hand, girls may be disadvantaged in such environments because individual personality traits may be inhibited from emerging over and above the gender-group related expectations. It is possible can, then, to hypothesise that teacher judgment accuracy might depend on classroom variables, because specific environments may lead teachers to process information about their students by relying on their own expectations.

Thus, I expect that classroom composition and school track moderate the extent of GGG, specifically:

H3a, Structural hypothesis: Smaller classrooms put conditions in place that reduce the GGG.

H3b, Composition hypothesis: In high SES classrooms and in those with a higher share of female students, the GGG is smaller.

H3c, Tracking hypothesis: Academic tracks are associated to a smaller GGG.

Features of the Italian Educational System

In the current Italian educational system, children enter schools at the age of 6, and school is mandatory until they are 16 years old. They attend eight years of comprehensive education, divided into five years in primary education, plus three years in lower secondary education. Primary school is commonly preceded by kindergarten, which lasts up to three years and is non-compulsory. In Italy's public schools, which in 2019 accounted for 92% of the total number[12], curricula are state-mandated and therefore similar across schools. After taking a national examination at the end of 8th grade, students are asked to make their first momentous choice among different educational programs, in order to attend upper-secondary schools.

There are several types of upper-secondary schools, but they can be broadly classified into three tracks: the academic track (*licei*), the technical track (*istituti tecnici*), and the vocational track (*istituti professionali*). The three main tracks have different curricula, subjects, educational purposes, and levels of heterogeneous prestige (Contini

[12] Calculation based on www.dati.Istat (Italian national statistical office), 2019 data.

and Triventi 2016). In general, attending lyceums leads to university, whereas technical and vocational schools combine a general with a vocational education, and are usually considered to be less demanding.

There are significant differences by student gender in specific high school track enrolment. On average, more than 60% of students enrolled in lyceums are female, with around 70% of these enrolled in classical lyceums, but less than 50% enrolled in scientific/applied science lyceums, and up to 28% enrolled in scientific lyceum with sport curricula. Female students comprise the vast majority in linguistic lyceums (80%), humanities and social science lyceums (89%), and artistic lyceums (71%). In contrast, technical schools and vocational schools host larger proportions of boys; their respective student populations are approximately 70% and 57% male (Miur 2017).

After five years of upper-secondary schooling, the student population takes a national test (*esame di maturità*). Neither the final test score nor the type of chosen school is binding for continuing to university, even if several tertiary degree programs require ability-based entry tests and accept only a limited number of students.

Concerning teacher evaluations, the Italian Ministry of Education (Miur[13]) offers an evaluation regulation (*regolamento di valutazione*[14]) that includes precise guidelines for how teachers are expected to grade their students. These guidelines differ according to education level. In primary education, grades are expressed as a descriptive evaluation for each subject, which reflects four levels of learning: 'advanced', 'intermediate', 'basic', and 'in the process of first acquisition'. In secondary schools, grades are instead

[13] Miur: Ministero dell'Istruzione, dell' Università e della Ricerca, divided in 2020 into two different ministries: Ministero dell' Istruzione (Ministry of Education) and Ministero dell' Università e della Ricerca (Ministry of University and Research).

[14] https://www.miur.gov.it/valutazione.

expressed as numbers from 1 to 10, where 6 is considered the passing mark and 10 is the highest score. During the academic years, students receive two report cards: one around February (*primo quadrimestre*) and one around June, at the end of the school year (*secondo quadrimestre*). Each subject has a distinctive grade: if at the end of the school year, students have not passed a given subject (meaning their subject grade is below 6), they need to take an exam on that subject before starting the new academic year. If the final report card contains three or more subjects below 6, the student needs to repeat their entire academic year. In the report card, grades are calculated as the average of results for several examinations and tests that students were required to take during the school year. Such examinations can vary according to the subject and to teacher preferences, but in most cases, they involve both oral and written testing.

Grading practices are also delineated in the formal educational agreement of co-responsibility (*patto educativo di corresponsabilità*[15]), also overseen by Miur. This document, which must be signed by both parents and students at the beginning of lower-secondary school, lists the principles and behaviours that schools, families, and students are expected to share and agree to respect. This emphasises the idea that grades are only one component of students' evaluation, that which also includes the context, students' progress throughout the school year, analysis of their overall decline and progress patterns, and the school situation as a whole.

Despite the specificity of both regulatory documents (*regolamento di valutazione* and *patto educativo di corresponsabilità*), the extent to which these policies are implemented by schools and teachers remains unclear. Indeed, they are conceived as mere guidelines; therefore, each school can determine its own evaluation criteria, which are

[15] https://www.miur.gov.it/web/guest/patto-educativo-corresponsabilita.

usually documented in school regulations. According to each school's specific regulations, teachers may have greater or lesser autonomy in deciding both the types of tests students must take, and their own grading practices.

Research Design

Data

The empirical analysis is based on data collected by the Italian National Institute for the Evaluation of the Education System (INVALSI) within the National Evaluation System (SNV). The main mission of INVALSI is to perform periodic, systematic assessments of students' knowledge skills, of the quality of national educational institutions, and of the quality of vocational training. The INVALSI-SNV dataset provides a variety of information on the entire population of students enrolled in specific grades and academic years; data are gathered via administrative sources and student questionnaires. In particular, these data contain information for both teacher assessments of student abilities (teachers' grades) and student scores on standardized tests in Language and Mathematics (INVALSI test scores). An important advantage of this data source is that information on teachers' grades comes directly from the schools, and is not self-reported by students, which notably increases reliability.

In 2012, INVALSI handed out for the first time a CAWI[16] questionnaire, which was addressed to a random sample of Language and Mathematics teachers who worked in specific grade levels. The questionnaire gathered information on both teachers' socio-demographic characteristics and their teaching habits: their professional profiles and

[16] CAWI: Computer Assisted Web Interviewing.

training, attitudes towards teaching, relationships with colleagues, and teaching practices. Thus, in the INVALSI-SNV dataset, it is possible to use student and class identifiers to link information about students with that of their teachers in the Language and Mathematics subjects. The final sample utilised in this study includes students in 10th grade matched with their teachers in the academic year 2015–16, pooled with students in 10th grade matched with their teachers in the academic year 2016–17. Our analytical sample includes 38,957 students with valid information for the variables included in the analysis.

Variables

Our dependent variables are teachers' grades in Mathematics and Language. Grade scoring ranges from 1 to 10, where 6 is the passing mark and 10 is the highest mark; grades are constructed as the average of oral and written performance marks. INVALSI collects information about grades from the midterm report card (around February), so information about teachers' grades at the end of the academic year is not available from their dataset. However, the INVALSI test is usually administered around March or April, making the two different kinds of assessment relatively close in time. This limits the risk that differences between the two indicators, teachers' grades and INVALSI test scores, would be substantially affected by differential learning across genders.

Our main independent variable is student gender, recoded as 0 for a male student and as 1 for a female student. The main control variable is the student score on the INVALSI standardized test on Language and Mathematics. Through the INVALSI test score, we have information about the subject-specific competences of students, measured by applying Item Response Theory (IRT) to students' answers (see Lord 2012 for further information on IRT). This information provides a measure of student ability that is independent of teacher assessments, since it is blindly evaluated and therefore is considered to be ‘unbiased’ (Borghans et al. 2016). INVALSI test scores have a mean of 200 and a standard deviation of 40. Following Triventi (2020), the argument is that INVALSI test scores are a good proxy of subject-specific competence compared to other indicators (such as the PISA test score), because they measure curriculum-related competences in the Italian educational system, and are therefore more directly comparable, in terms of knowledge, with what is expected to be assessed via school marks. Indeed, the main scope of INVALSI is assessing students’ knowledge according to what national legislation expects for that specific grade. By definition, INVALSI tests do not aim to assess additional students’ abilities such as communication skills, emotional and relational skills, participation and engagement. In fact, these dimensions are expected to be embedded in teacher assessed grades[17]. INVALSI test score results are also adjusted in order to reduce the risk of cheating during test administration (i.e., INVALSI 2018).

For other control variables at the student level, I use quarter of birth, migration background, geographical area, regularity in studies, attendance in kindergarten, an index of Economic and Social-Cultural Status (ESCS)[18], and a control for the academic year. These variables are included because previous studies suggest they are associated with teachers’ grades.

[17] For more information about INVALSI test score: www.invalsiopen.it.

[18] Index provided from INVALSI that measures students’ economic, social and cultural status. It is a synthesis of three indicators: 1. Parental occupational status; 2. Parental level of education; 3. Possession of specific material assets.

As moderator variables at the second level, I introduced some teacher characteristics expected to affect the size of the GGG: gender, age[19], within-school seniority[20], and contract type. Concerning classroom composition, I consider the percentage of female students, the percentage of students with medium-high and higher ESCS[21], and the classroom size.

As moderator variables at the school level, I control for the type of school attended by students (lyceum, technical school, and vocational school). Table 3.1 provides information about the recoding.

[19] We tested the model fit introducing also age squared. Both Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) suggest that age squared does not improve the model fit.

[20] The expression "within-school seniority" refers to the length of teacher's experience within the school.

[21] Medium-high and higher ESCS students correspond to the 3rd and 4th quartile of ESCS distribution

Table 3.1: Description of the variables used in the analysis.

Variable	Description
Dependent variable	
Teacher grade in Mathematics Teacher grade in Language	Teacher grade in mid-term school report (*pagella*) in February. Average between grade in oral exam and grade in written exam. Scale 1-10 (where 6 is the passing mark)
Independent variable	
Student gender	Recoded as (0) if male; (1) if female
Student control variables	
INVALSI test score in numeracy INVALSI test score in literacy	Scores obtained by students in the INVALSI standardized tests. The scores have mean 200 and standard deviation 40. For the analysis, I standardized to have a mean of 0 and a standard deviation of 1. I use scores adjusted for potential cheating.
Quarter of birth	Recoded as (0) (January, February, March); (1) (April, May, June); (2) (July, August, September); (3) (October, November, Dicember)
Migration background	Recoded as (0) if students are native born; (1) if students are II generation immigrants; (2) if students are I generation immigrants
Geographical area	Recoded as (0) North West; (1) North East; (2) Centre; (3) South; (5) Isles (as defined by ISTAT, national istitute of statistics)
ESCS	Index provided from INVALSI that measures students' economic, social and cultural status. It is a synthesis of three indicators: 1. Parental occupational status (HISEI); 2. Parental level of education (PARED); 3. Possession of specific material assets (HOMEPOS), I recoded the variable in (0) Lower (1st quartile); (1) Medium-low (2nd quartile); (2) Medium-high (3rd quartile) (3) Higher (4rt quartile)
Regularity in studies	Recoded as (0) if students are regular; (1) if students are early starters (*primina*); (2) if students are late starters (including rejection)
Attendance in kindergarten	Recoded as (0) if students did not attend kindergarten; (1) if students attended kindergarten
Academic year	Recoded as (0) if students were in 10th grade in the A.Y. 2015/2016; (1) if students were in 10th grade in the A.Y. 2016/2017
Language and mathematics teacher variables	
Gender	Recoded as (0) if teachers are male; (1) if teachers are female
Age	Continuous variable, scale 28-77
Seniority in school	Recoded as (0) if teachers work in the school from 1 year or less; (1) from 2 to 3 years; (2) from 4 to 5 years; (3) from more than 5 years
Type of contract	Recoded as (0) if teachers have a fixed-term contract; (1) if teachers have a permanent contract
Classroom and school variables	
Class size	Number of students in the classroom, scale 5-30
% Female students	Percentage of female students in the classroom, scale 0-100
% High-background students	Percentage of students with medium-high (3rd quartile) or higher (4th quartile) ESCS in the classroom, starting from our recoding of ESCS at the student level, scale 0-100.
Type of school	Type of school attended by students, recoded as (0) Vocational schools; (1) Technical schools; (2) Lyceums

Analytical Strategy

The objective of the analysis is twofold. The first goal is to test our first hypothesis, establishing to what extent female students are graded more generously than boys by their teachers. A common strategy used to assess whether teachers grade better one group compared to another with the same academic abilities and competences is the grade equation model (for examples, see Triventi 2020; Kiss 2013; and Hinnerich, Höglin and Johannesson 2011). The basic form of the grade equation model is a regression, in which a non-blind measure of student performance (in setting, teacher's grade) is expressed as a function of the variable identifying the group of interest (student gender) plus a blind measure of student ability (INVALSI test score) and a series of control variables.

The measure of teacher gender bias, or teacher grading premium, is therefore constructed by considering the average marks for boys and girls in Mathematics and Language in 'non-blind' classroom exams, and the respective mean scores of a 'blind' national exam (INVALSI test) that is marked anonymously.

Teacher grading premium or penalty for females is assessed relying on two different models for each subject (s), Language and Mathematics:

$$Grades_{i_s} = \alpha + \delta(female_{i_s}) + \beta(testscore_{i_s}) + \varepsilon_{i_s} \qquad (1)$$

In this basic form of the grade equation model (eq. 1), δ represents the extent of teacher grading premium or penalty towards female students: if $\delta > 0$ then females are over-assessed, and if $\delta < 0$ then females are under-assessed compared to males. Test scores are considered in this framework as yardsticks against which teachers' marks can be compared[22]. The grade equation model allows us to flexibly handle the different measurement scales of the two competence measures (Dardanoni, Modica and Pennisi 2009; Hinnerich, Höglin and Johannesson 2011; Kiss 2013). In order to properly measure gender grading bias, I rely on two grading equation models embedded in hierarchical linear regression analysis, in which students comprise the first level, classrooms and teachers comprise the second level, and schools comprise the third level. The two main advantages of multilevel modelling over OLS regressions are: the possibility of considering the natural nesting of the data, thus solving the problem of dependency of observations; and the likelihood of providing correct standard errors, thus avoiding underestimating them, which would lead to incorrect inferences or interpretations.

In the first step of our analysis, I estimate a random intercept model for each of the two subjects (s), in which the average grade is allowed to vary by classroom, but the slope of the regression line is assumed to be fixed across schools and classrooms:

$$Grades_{ijk_s} = \alpha + \delta(female_{ijk_s}) + \beta_1(testscore_{ijk_s}) + \beta_2(I_{ijk_s}) + \beta_3(T_{jk_s}) + \beta_4(C_{jk_s}) + \beta_5(Z_{k_s}) + u_{0j_s} + u_{0jk_s} + \varepsilon_{ijk_s} \qquad (2)$$

[22] Another possibility would be to assess the teacher grading premium with a difference-in difference design (Di Liberto, Casula and Pau 2021). Teacher grading premium can be understood as the difference between girls and boys, gathered via the difference between teacher grade ('non-blind' measure) and the standardized test score ('blind' and supposedly unbiased measure) (see Lavy 2008). Consequently, teacher grading mismatch would be defined as the average gap between for female students, minus the same gap for male students, as prior researchers have done to estimate gender discrimination (Falch and Naper 2013; Goldin and Rouse 2000). However, in the current context, this would also be feasible by standardizing the different scores in order to compare them.

These models follow the usual multilevel structure, where students (i) are nested in classrooms (j), which are nested in schools (k), and where:

- $Grades_{ijk_s}$ is the dependent variable that measures teacher assessment, with scores ranging from 1 to 10;
- δ represents the extent of teacher grading premium favouring females;
- $\beta_1(testscore_{ijk_s})$ represents the main control variable at the individual level: the INVALSI test score, with its associated regression coefficient;
- $\beta_2(I_{ijk_s})$ is a vector of control variables at the individual level, with its associated regression coefficients;
- $\beta_3(T_{jk_s})$ is a vector representing teacher characteristics at the classroom level, with its associated regression coefficients;
- $\beta_4(C_{jk_s})$ is a vector representing classroom composition variables, with its associated regression coefficients; and
- $\beta_5(Z_{k_s})$ is the type of school attended, with its associated regression coefficients.

The second step of the analysis is devoted to testing our second and third hypotheses, aimed at understanding which characteristics of teachers and classrooms are likely to enlarge or reduce the GGG. To do so, I estimate a series of random coefficient multilevel models, in which the GGG is allowed to vary across classrooms, and in which cross-level interactions are introduced to assess whether the GGG varies across: i) teacher gender (resemblance hypothesis); ii) teacher characteristics such as age, within-school seniority, and contract type (experience hypothesis); iii) classroom size (structural hypothesis); iv) percentage of female students and percentage of students from higher social background (composition hypothesis); and v) school track (tracking hypothesis).

The multilevel random slope regressions are estimated using the following model specifications, with cross-level interactions between gender and, respectively, teacher characteristics (eq. 3), classroom composition (eq. 4), and school track (eq. 5):

$$\begin{aligned}Grades_{ijk_s} = {} & \alpha + \delta(female_{ijk_s}) + \beta_1(testscore_{ijk_s}) + \beta_2(I_{ijk_s}) + \beta_3(T_{jk_s}) + \beta_4(C_{jk_s}) + \\ & \beta_5(Z_{k_s}) + \delta(female_{ijk_s})u_{1jk_s} + \boldsymbol{\beta_6}(\boldsymbol{female_{ijk_s}}) * (\boldsymbol{T_{jk_s}}) + u_{0j_s} + u_{0jk_s} + \varepsilon_{ijk_s}\end{aligned} \tag{3}$$

$$\begin{aligned}Grades_{ijk_s} = {} & \alpha + \delta(female_{ijk_s}) + \beta_1(testscore_{ijk_s}) + \beta_2(I_{ijk_s}) + \beta_3(T_{jk_s}) + \beta_4(C_{jk_s}) + \\ & \beta_5(Z_{k_s}) + \delta(female_{ijk_s})u_{1jk_s} + \boldsymbol{\beta_6}(\boldsymbol{female_{ijk_s}}) * (\boldsymbol{C_{jk_s}}) + u_{0j_s} + u_{0jk_s} + \varepsilon_{ijk_s}\end{aligned} \tag{4}$$

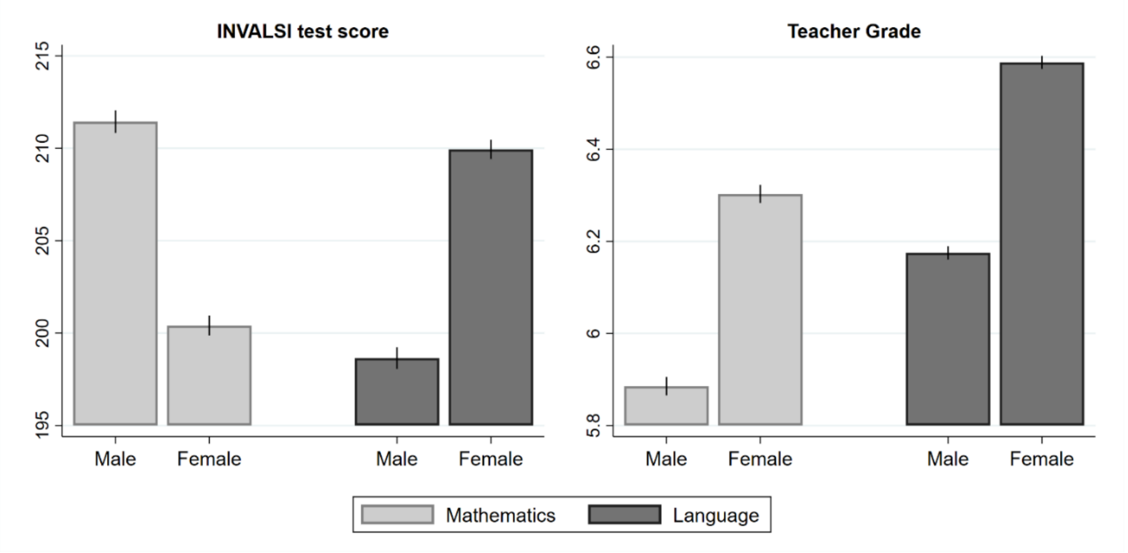
$$\begin{aligned}Grades_{ijk_s} = {} & \alpha + \delta(female_{ijk_s}) + \beta_1(testscore_{ijk_s}) + \beta_2(I_{ijk_s}) + \beta_3(T_{jk_s}) + \beta_4(C_{jk_s}) + \\ & \beta_5(Z_{k_s}) + \delta(female_{ijk_s})u_{1jk_s} + \boldsymbol{\beta_6}(\boldsymbol{female_{ijk_s}}) * (\boldsymbol{Z_{k_s}}) + u_{0j_s} + u_{0jk_s} + \varepsilon_{ijk_s}\end{aligned} \tag{5}$$

Results

Gender Grading Mismatch

As a first step, in Figure 3.1, I examine the average test scores and marks in Language and Mathematics among boys and girls. Analysing standardized INVALSI test scores makes evident that boys outperform girls in Mathematics, while girls perform significantly better in Language. This result is not surprising, compared with the gender gap in achievement across the OECD countries (OECD 2009, 2014; Fryer and Levitt 2010).

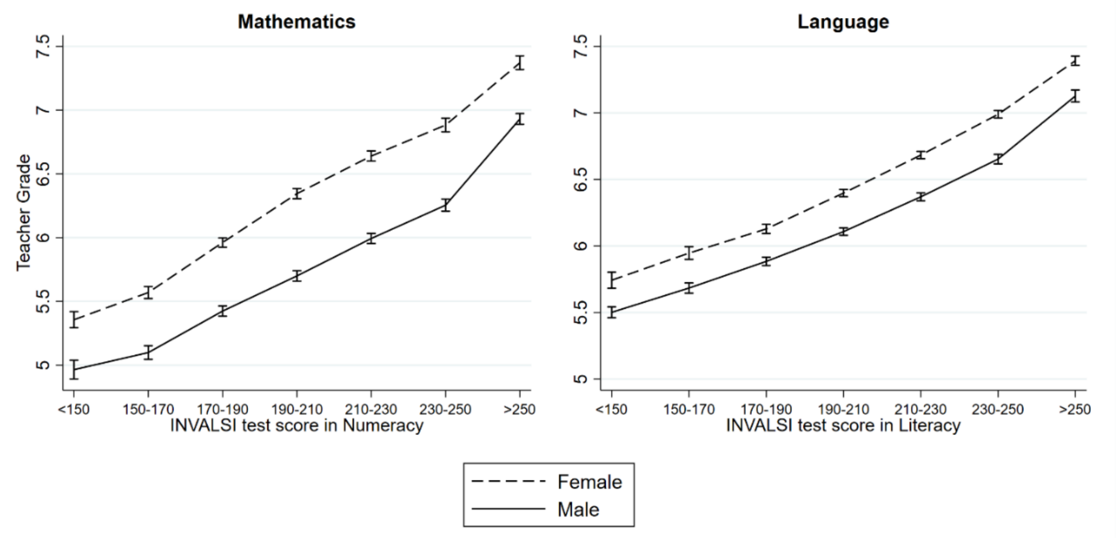
Figure 3.1: Average INVALSI test score in Numeracy and in Literacy for male and female students (left) and average teacher grades in Mathematics and Language for male and female students (right); 95% confidence intervals; N = 38,975



In both subjects, the difference in the average teacher grade between male and female students is approximately 0.4 points. The average grade in Language is around 6.2 for boys, while it is 6.6 for girls. In Mathematics, female students have on average a grade equal to 6.3, while boys are on average under the passing mark, performing at 5.9.

In Figure 3.2, I report the average grades in Language and Mathematics across boys and girls, along the distribution of the subject-specific competence level. The figure clearly shows a female grading premium: girls receive, on average, higher grades than boys at the same level of subject-specific competence. Additionally, it appears that such female advantage occurs across the entire distribution of students' subject-specific competences (INVALSI score). One might expect the gap to be larger in the middle of the competence distribution, because students around the mean, with average competences, may be graded more in accordance with their gender-related non cognitive skills and behaviours, rather than according to their actual competences.

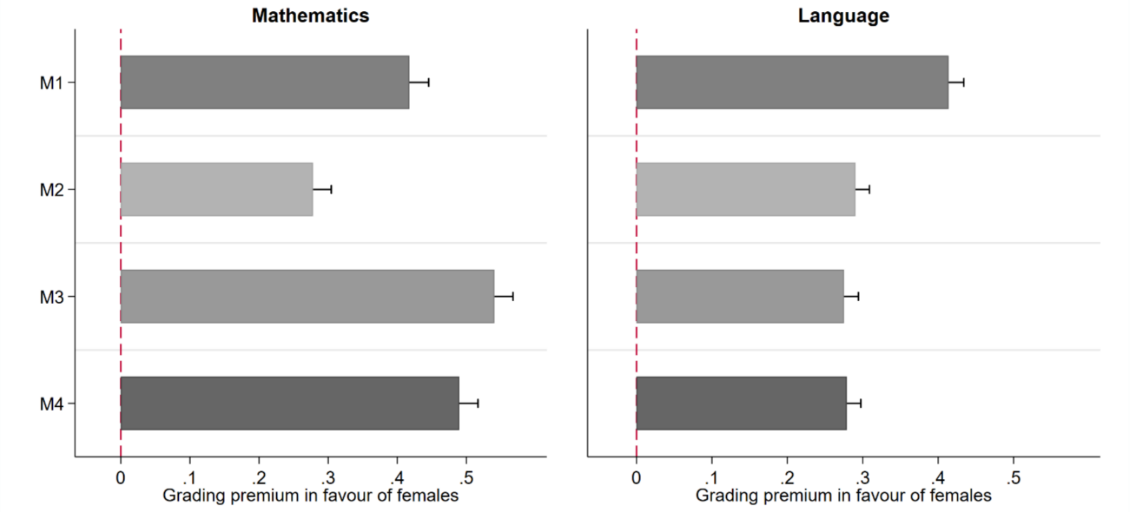
Figure 3.2: Average teacher grade for male and female students in Mathematics (left) and Language (right) for different intervals of INVALSI test score distribution; 95% confidence intervals; N = 38,975.



Indeed, students at the extremes of the distribution – those who perform very poorly at one end, or very well at the other end – are thought to be evaluated in relation to their competences, as the stronger signal. However, Figure 3.2 shows clearly that the gender gap in teacher evaluation does not vary according to the level of competence. This gap is surprisingly constant in Language, in which girls are evaluated about 0.4 points higher than boys at each increment of the INVALSI score distribution. In Mathematics, the gap appears larger in the middle of the distribution, and slightly reduces for the top-performing students. Despite that, the difference in grades between boys and girls, even the top performing ones, is not negligible and statistically significant.

We next turn to the estimation of the GGG via regression models, as Figure 3.3 presents. The first three models are estimated via OLS linear regression. Model 1 represents the simple average gender gap in grades; model 2 adjusts for subject-specific competences, as measured by INVALSI test score; model 3 adjusts for INVALSI test score and individual level control variables; and model 4 is a three-level random intercept model with students nested in classrooms nested in schools, with the same independent variables as model 3.

Figure 3.3: Teacher gap in grading in favor of female students in Mathematics (left) and in Language (right). Coefficients derived from four linear prediction models of teacher grade for female and male students (male = reference category); 95% confidence interval; N = 38,957.



Note: Model 1 represents the simple average gap in grade; model 2 controls for subject-specific competences (INVALSI test score); model 3 controls for INVALSI test score and individual level controls (student quarter of birth, migration background, geographical area, ESCS, regularity in studies, attendance to kindergarten, academic year); in model 4 linear predictions are derived from a hierarchical model with students nested in classrooms nested in schools, with the same controls as model 3.

Comparing the first two models, girls evidently obtain better grades than boys in both subjects; once we adjust for their subject-specific competence, the average differences are only slightly reduced. This means that only to a minor extent is the advantage of girls in grades due to their superior subject-specific competences, a result that is further confirmed in models 3 and 4. Comparing model 1 with model 4, the gap between boys and girls in teachers' grades appears to be slightly reduced in Language, even if it remains substantially relevant and statistically significant, whereas in Mathematics the GGG enlarges.

Teacher Characteristics

Having established to what extent females are graded more generously than boys by their teachers, the next goal is understanding whether the sign and magnitude of this pattern is related to specific teacher characteristics: gender, age, within-school seniority, and contract type. This is achieved by relying on random slope models, in which student gender is allowed to have a different effect on teachers' grades across classrooms, and which include a cross-level interaction between student gender and teacher characteristics[23]. Overall, our results indicate that GGG in favour of females does not change according to teacher characteristics, either in Mathematics or in Language. In other words, the female grading premium is always present, irrespective of teachers' individual characteristics and practices.

Figure 3.4 shows the predictive margins and average marginal effects of cross-level interactions between student gender and teacher characteristics in Mathematics. The average marginal effects of cross-level interactions between student gender and teacher gender shows that both male and female maths teachers, on average, attribute higher grades to girls than to boys who have the same subject-specific competence. This suggests that male teachers are not (unconsciously) likely to better reward male students, and female teachers do not penalize more male students, as suggested by the same-sex resemblance hypothesis.

[23] Before introducing teachers and classrooms predictors, models with random slopes without predictors at the higher levels are performed, in order to control whether gender grading gap varies across teachers and classrooms characteristics even without including higher level predictors. Figure A3.1 (in the appendix) suggests that the relationship between student gender and student grades does not vary substantively across classrooms without including higher level predictors.

Figure 3.4: Predictive margins and average marginal effects of cross-level interactions between student gender and teacher characteristics in Mathematics (gender, age, within-school seniority, contract type); coefficients derived from multilevel model; 95% confidence intervals; N = 38,957.

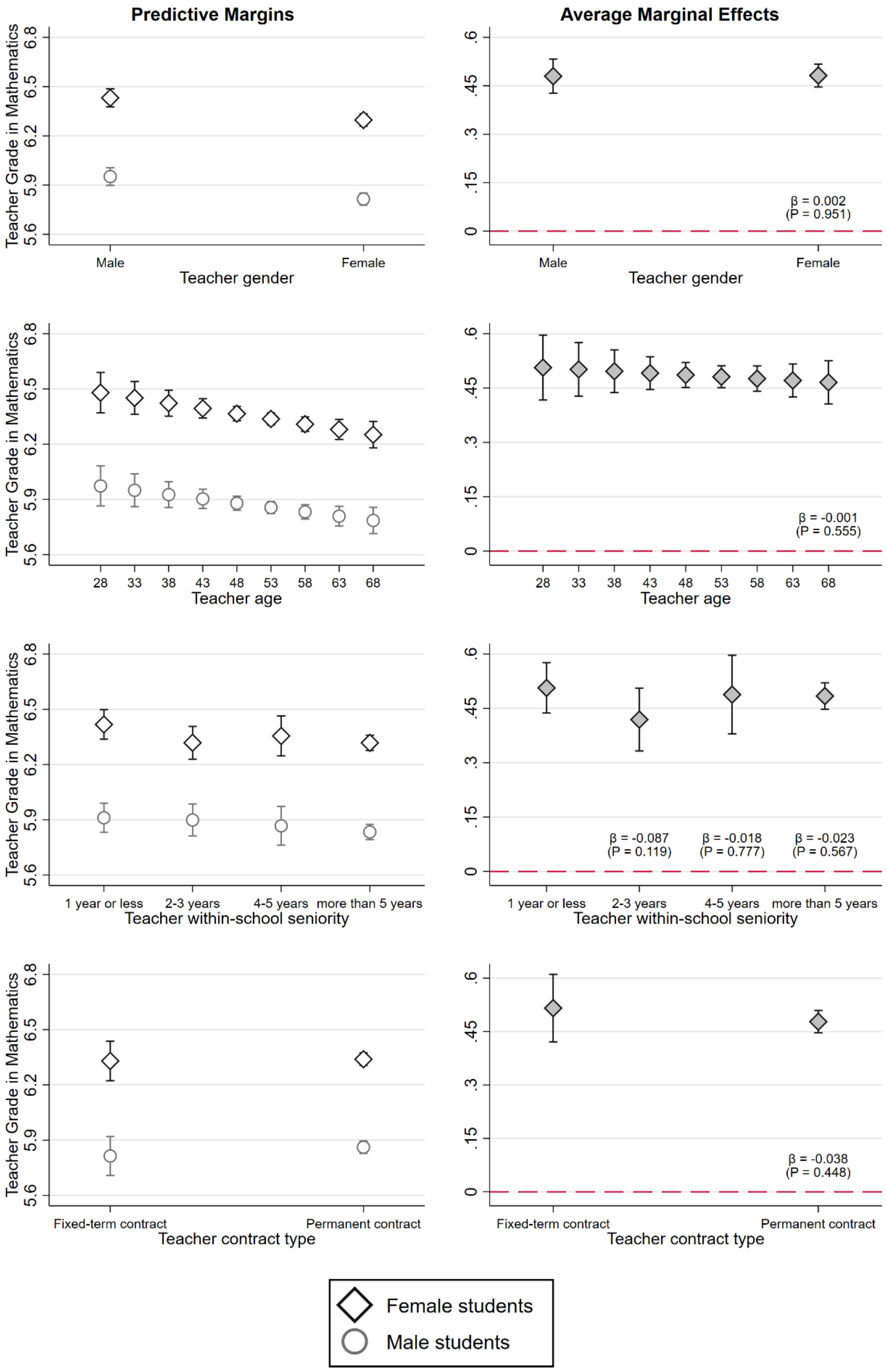
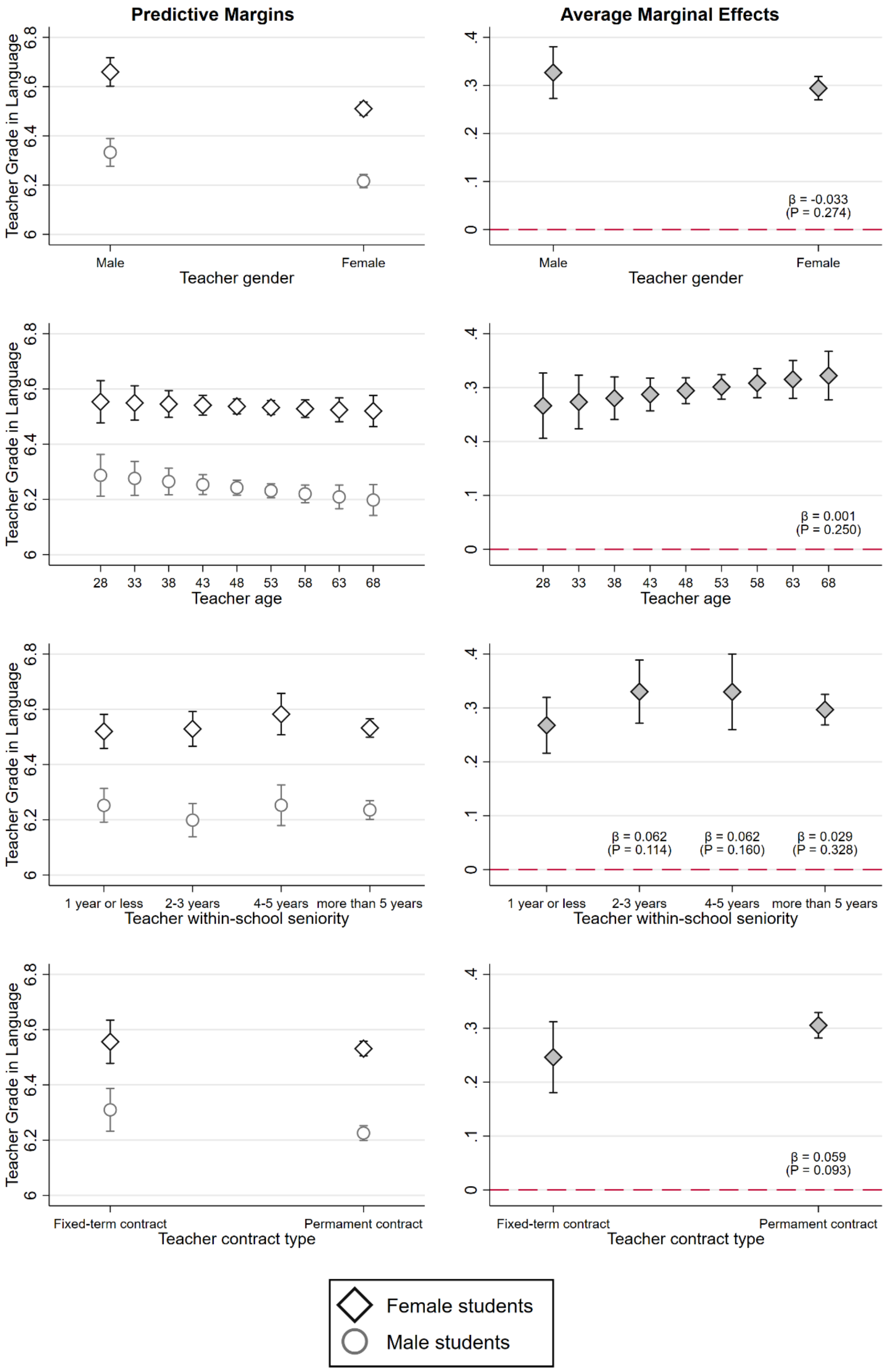


Figure 3.5: Predictive margins and average marginal effects of cross-level interactions between student gender and teacher characteristics in Language (gender, age, within-school seniority, contract type); coefficients derived from multilevel model; 95% confidence intervals; N = 38,957.



In Mathematics, while older teachers tend to give on average lower grades, having an older, and therefore more experienced, teacher does not affect the levels at which female and male students are graded. Examining within-school seniority yields similar results: maths teachers who have more than five years of seniority in school are more severe in grading, compared with teachers who have one year or less of seniority, but the difference in grading between girls and boys is similar irrespective of the teacher's length of experience within the school. Therefore, I cannot conclude that teachers with more seniority, who are likely to better know their students, are better able to provide more homogeneous grading for boys and girls who have the same mathematics competences. In addition, a teacher's type of contract does not affect either the extent of GGG, or the average grade of students.

Figure 3.5 shows a similar finding for the interaction between student gender and teacher characteristics in Language.

Classroom Composition and School Type

Regarding classroom composition, an increase in the size of the classroom corresponds to a slight decrease in the average grade. However, only for Mathematics the number of students per class significantly moderates the extent of the GGG. Indeed, the GGG is bigger when the classroom size is larger. As Figure 3.6 shows, in Mathematics classes, boys are more penalized than girls in terms of grading when the number of students per classroom increases.

Interestingly, an increase of the percentage of female students in the classroom corresponds to an increase in the overall average grade for Mathematics, but also to a slight decrease in the overall average grade in Language. However, the share of girls in the classroom does not significantly moderate the GGG. Finally, a higher share of students with medium-high and higher ESCS is associated with a lower average grade only in Mathematics, but it does not moderate the extent of GGG either in mathematics or in Language.

Overall, we can conclude that classroom composition features do not significantly contribute to enlarging or reducing the gender grading gap, with the exception of class size for Mathematics.

For the type of school attended[24], results show that the average grade is lower in both lyceums and technical schools compared to vocational schools; this is true for both Language and Mathematics. Moreover, the type of school students attend seems to partially moderate the GGG, only in Mathematics; the gap is significantly larger in both technical schools and lyceums, compared with vocational schools. Therefore, in Mathematics, attending a technical school or lyceum rather than vocational school increases the gap in grading to the detriment of male students, who seem to be even more disadvantaged compared to their female counterparts. In contrast, regarding Language (Figure 3.7), attending a lyceum or a technical school rather than a vocational school does not affect the gap in grading between male and female students.

[24] It is important to note that these models also include controls for classroom composition. Therefore, our estimates regard GGG for female and male students in different types of schools that have similar classroom composition. Estimates without controls for classroom composition suggest similar patterns.

Figure 3.6: Predictive margins and average marginal effects of cross-level interactions between student gender and classroom and school characteristics in Mathematics (class size, % of female students, % of high-ESCS students, school track); coefficients derived from multilevel model;95% confidence intervals; N = 38,957.

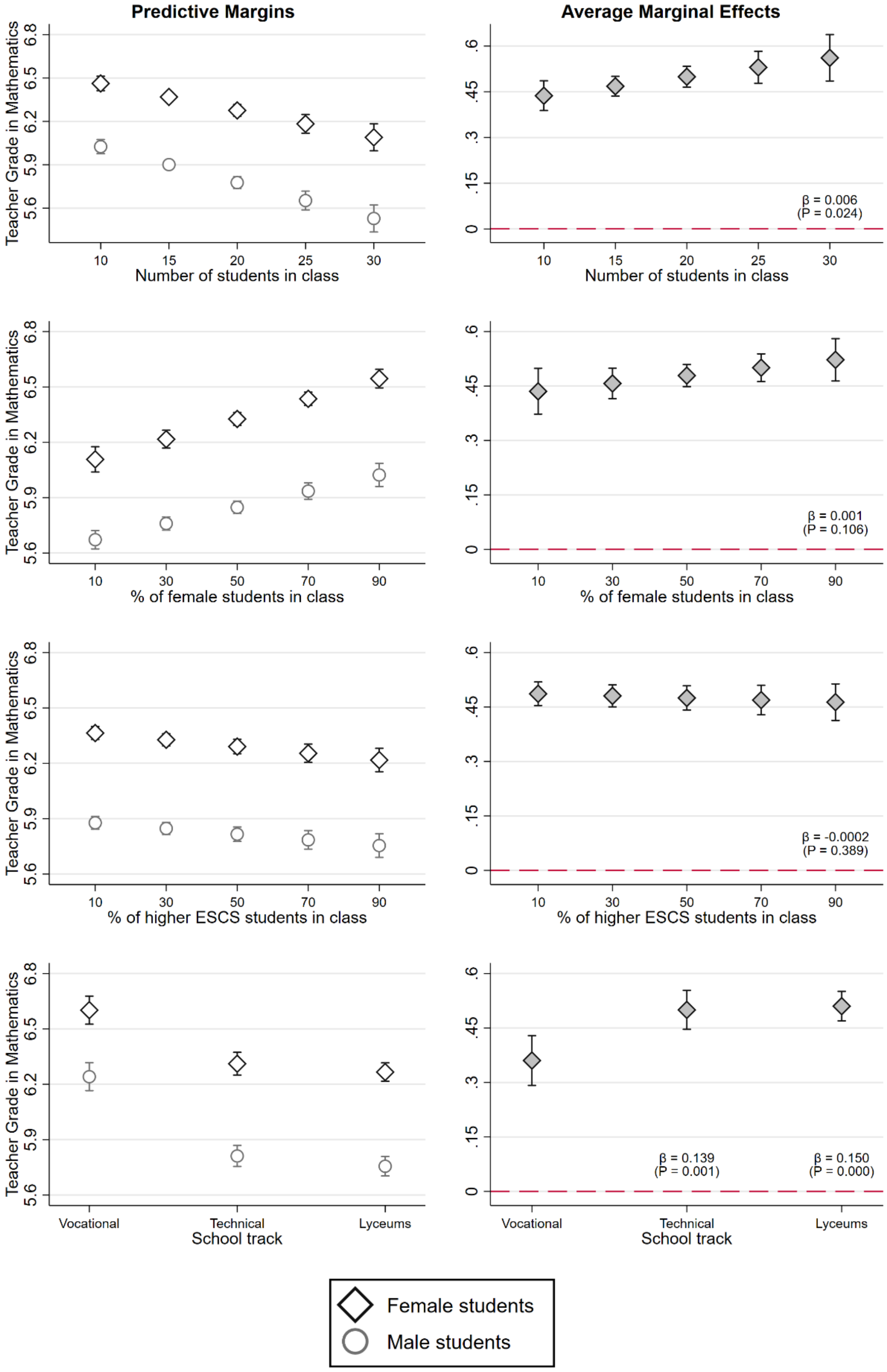
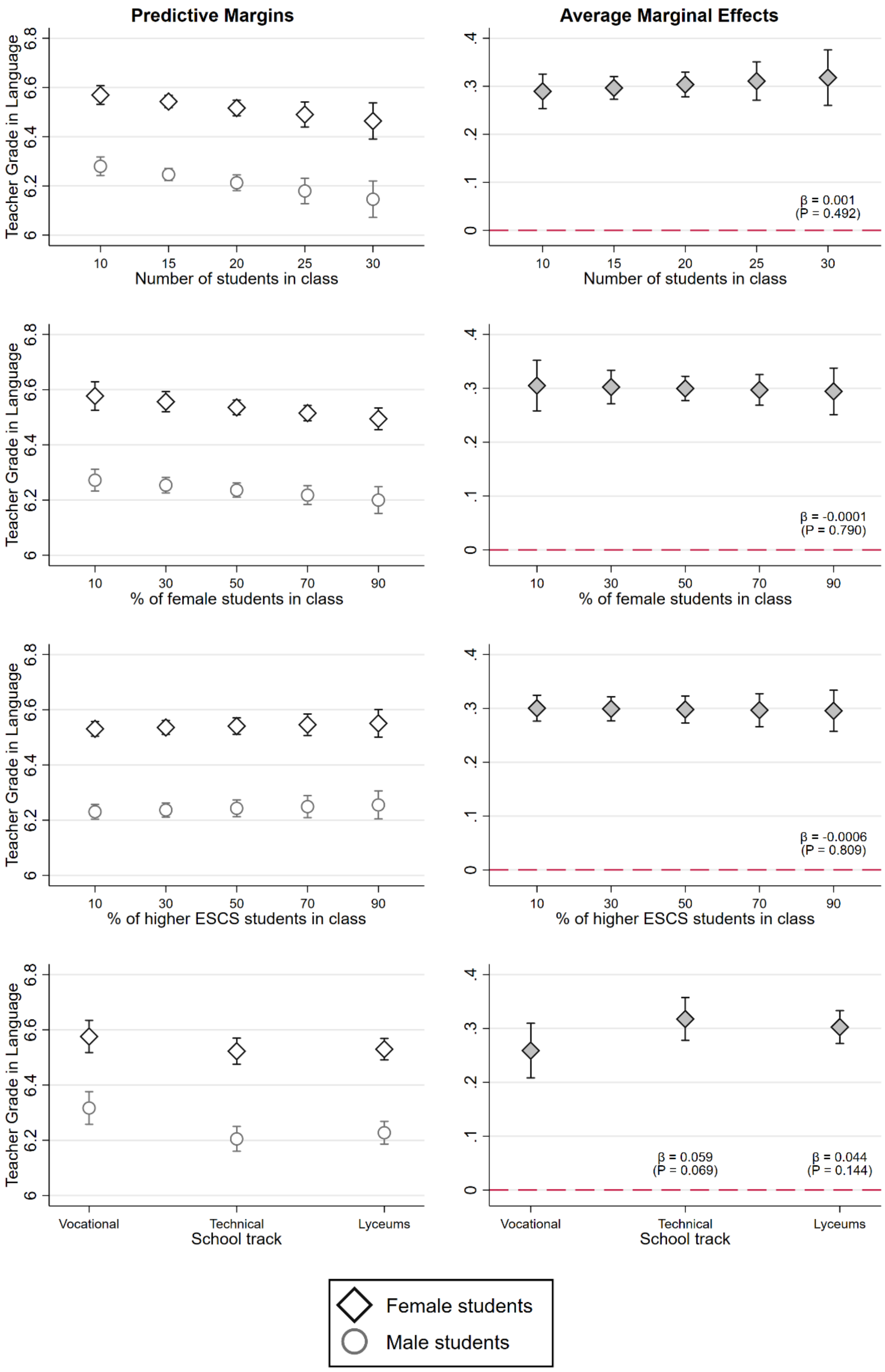


Figure 3.7: Predictive margins and average marginal effects of cross-level interactions between student gender and classroom and school characteristics in Language (class size, % of female students, % of high-ESCS students, school track); coefficients derived from multilevel model; 95% confidence intervals; N = 38,957.



Robustness Checks

Three additional analyses are proposed in order to control for the robustness of the results (see online Appendices A3.5 for Mathematics and A3.6 for Language). The first analysis includes fixed-effect terms at the second level of analysis, in order to account for possible omitted-variable bias due to unobserved heterogeneity at the classroom level. Controlling for all possible confounders, including teacher characteristics, at the second and third levels, coefficients are almost identical, and they do not change the substantial interpretation of results. Therefore, our set of control variables at the classroom and school levels are good predictors of teacher grade in the two subjects.

The second additional analysis includes inverse probability weighting (IPW), in order to control for possible selection bias in our analytical sample. The final sample for Mathematics includes 42,707 students, while the final sample for Language includes 41,880 students. In order to compare coefficients across the two subject areas, I selected the smallest sample across both, which includes students who perform on both the literacy and numeracy tests. This implies a possible exclusion of students who decided to skip the exam for one of the two subjects, leading to biased estimates. To resolve this issue, I constructed different IPWs for Language and for Mathematics, based on the probability of the student being absent on the day of the test, using socio-demographic student variables as predictors, and including teacher grades. Results of the chosen models are robust.

The third additional analysis controls for measurement error in the INVALSI test score; in comparison with teacher grade, which is a repeated measure over the course of the academic year, the INVALSI test score may be subject to measurement error resulting from a variety of factors, such as a student's emotional or mental state while taking the test, or the conditions under which the test was administered. For this reason, GGG may be over- or under-assessed. Specifically, regarding Mathematics, as boys outperform girls in the INVALSI test score, the grading gap between the two groups may be under-assessed. In contrast, in Language, female students outperform male students in the INVALSI test score, resulting in a possibly over-assessed gap. To assess the magnitude of the bias in the estimates of subject-specific competences, following Bound, Brown and Mathiowetz (2011), I adopt an instrumental variable (IV) estimation approach. In this framework, the IV is constituted by a lagged student's subject-specific competence score. For each 10th-grade student in the analytical sample, the IV is represented by their individual test score from the 8th grade (in the academic years 2013–14 and 2015–16, respectively). This linkage is made possible by the INVALSI-SNV dataset structure, in which the SIDI code, a student unique identifier, allows the matching of students with their INVALSI test score results at previous grade levels. This approach enables correction of the extent of the teacher grading premium favouring female students. However, a considerable problem regarding this specification concerns the linkage with the SIDI codes, which implies a loss of cases and missing values due to school non-reporting, ID misclassifications, or students who did not pass the school exams and are thus not attending the originally planned academic year. Further, it implies a possible self-selection in the dataset of high-performing students which can affect the final estimates, and lead to a considerably smaller sample. Therefore, our preferred model remains the first. As expected, using the INVALSI test score from 8th grade as the instrument, the gap in Mathematics slightly increases, while in Language it slightly reduces; however, the overall findings are confirmed.

Conclusions and Discussion

This chapter has aimed to explore the role of teachers in determining gender differences among academic outcomes in Italian upper-secondary schools. The goal is providing a measure of the gender-based grading mismatch – the difference in grade between male students and female students who have identical subject-specific competences. Previous research shown that girls are graded more generously than boys, even when they have the same level of competence, as measured via standardized tests (Lavy 2008; Lindahl 2007; Angelo 2014; Terrier 2015; Enzi 2015; Kiss 2013). The current chapter addresses this gap, focusing on the Italian case, by demonstrating the existence of a gender-based grading mismatch that favours girls. Additionally, teacher premium in grading for girls versus boys appears to be different across two subjects, Language and Mathematics.

Focusing on grades in Language, for which subject females outperform males on average in the INVALSI test, it appears that teacher grading mismatch enlarges the GGG. In contrast, for Mathematics, in which boys outperform girls on the INVALSI test, the pattern is different. Teacher grading mismatch evens the gap, to the point that female students are always advantaged in classroom grades, despite their lower subject-specific competences. More specifically, girls are regularly assessed at 0.4 grade points higher than boys on average, in both Mathematics and Language. Whereas this estimate could be slightly overestimated for Language, the IV model specification suggests that in Mathematics the gap could be even larger. This may imply that teacher grading mismatch could be a significant discriminant in Mathematics, between scoring above or below the minimum passing grade. Therefore, I accept hypothesis H1: teachers are more likely to grade female students more generously than male students who have the same subject-specific competences.

A further, substantive contribution of the current study is the assessment of whether specific teachers' characteristics, classroom composition, and type of school attended affect the ways in which females and males are evaluated by their teachers. These characteristics have been selected on the basis of theoretical considerations and the results of prior research. This study's results suggest that overall, the teacher grading mismatch in favour of female students is systemic: teacher characteristics, classroom composition, and school type do not appear to have any considerable effect on decreasing the gender gap in grading. On the contrary, this study suggests that only classroom size and the type of school have a moderating role. For the Mathematics subject, some conditions such as classrooms with a large number of students, and classes within a technical or academic track rather than a vocational track, are associated with an increase in the gender grading mismatch. For Language instead, none of the considered features is significantly associated with the GGG.

Therefore, I can reject both the resemblance hypothesis and the experience hypothesis: teacher considered characteristics do not moderate the GGG. Regarding classroom characteristics, I can also reject the composition hypothesis, because the share of females and the share of higher-SES students do not affect the extent of the GGG.

However, I can partially accept the structural hypothesis, as a larger classroom size is significantly associated with an increase in the GGG, but only for Mathematics.

Finally, I can also partially accept the tracking hypothesis: the GGG is wider in both technical tracks and lyceums compared to vocational tracks, again only for Mathematics.

This (almost) systemic gender grading mismatch in favour of female students, which is surprisingly large especially in Mathematics, can be explained both methodologically and theoretically. Technically, the implemented models with fixed effects at the classroom level enable controlling for all the characteristics of classrooms and schools that might aid in explaining the GGG. However, this does not account for students' specific educational signals[25] that work beyond competences, such as behaviour in the classroom, participation, engagement, perseverance, and effort. Indeed, students' attitudes and behaviours in the classroom are relevant criteria for grades attribution, and they partially enter in teacher's evaluation, but they are irrelevant criteria for results on the INVALSI test. One related theoretical stream interprets gender grading mismatch as also being a function of students' observed behaviours. School and classroom environments might indeed be adapted to traditionally female behaviours (Lavy 2008). Female students might thus adopt such *actual* behaviours during class, including precision, order, modesty, and quietness, which go beyond the individuals' academic performance, but which teachers may highly reward in terms of grades. Indeed, the idea that teachers may be prone to favour 'girly' attitudes in classroom is corroborated by other Italian findings in studies examining earlier school grades (Di Liberto, Casula and Pau 2021). Conversely, teachers may be likely to associate such behaviours only with female students, because girls are traditionally thought of as possessing these traits. Consequently, teacher grading premium favouring females could also be related to teachers' *expectations* regarding their female students, rather than related to the actual behaviours of the latter during class.

Another theoretical explanation calls into play teacher overcompensation towards females. Girls are indeed often discussed in discrimination contexts, especially in speeches about gender differences in cognitive ability in dealing with the 'hard subjects'.

[25] INVALSI-SNV data does not collect and provide such information.

A possible explanation for the reason teachers are more generous in grading female students could be that teachers wish to avoid possible discrimination against girls as an ability-stigmatized group. Therefore, teachers may over-assess girls in the same way they sometimes over-assess non-native students, to avoid negative stereotyping (Alesina 2018). From another perspective, it is also possible that the over-assessment of female students' competences in Mathematics partly represents a sort of 'push' to encourage the weaker students (Terrier 2015). As a consequence, this may translate into a positive discrimination which favours girls.

A final interpretation of our results could stem from a limitation of our data. Indeed, our measure of student grading is derived from the midterm report card, rather than the final report card of the academic year. Teachers may consciously adopt specific grading practices at the midterm to trigger students' effort differently according to their gender, to in turn obtain the best possible performance by the end of the academic year. Specifically, teachers could perceive that male students' effort is more easily triggered by lower grades, which could encourage them to study harder to achieve a better grade, while female students' effort to achieve higher grades may be more easily triggered via encouragement.

Disentangling such mechanisms is beyond the scope of this chapter, however; this contribution to the body of research focuses on providing an estimate of the total effect of students' gender on teachers' grades. The main limitation of this chapter is the unavailability of indicators for students' behaviour and attitudes, which may be important mediators in explaining the teacher grading premium favouring females. Teacher assessment is indeed constituted by multiple different factors, also including students' non-cognitive indicators (Lipnevich et al. 2020); thus, it is not possible to entirely disentangle a pure gender discrimination bias from a broader grading gap related to students' attitudes and behaviours in class. Future studies should address the inclusion of students' gender-related behaviours and attitudes towards the subject matter. Moreover, a measure of teacher gender stereotypes is absent from this study, and this would be determinant to conclude that GGG is driven by teachers' discrimination against boys – or in favour of girls. Overall, it would be necessary to take into account unobserved heterogeneity before speculating about the meaning and the implications of such a grading mismatch.

In conclusion, the magnitude of the bias against male students in not negligible, and may have negative consequences. This is especially true regarding Mathematics, where a teacher penalty may translate into a failing grade, since the average teacher grade for boys falls right on the passing mark. Indeed, it has been hypothesized that boys' struggles in the Italian system might be partially driven by grading biases (Di Liberto, Casula and Pau 2021).

The interpretation of results using teacher grades rather than scores from standardized tests may also have important implications in terms of public policy, as teachers' grades have previously been found to be strong predictors of a variety of important life outcomes (Borghans et al. 2016).

Appendix Chapter 3

Table A3.1: Multilevel regression model on teacher grade in Mathematics; model 1: random intercept model with student-level explanatory variables; model 2: random intercept model adding classroom-level and school-level explanatory variables; model 3: random gender coefficient model with student-level, classroom-level and school-level explanatory variables. *** p<0.01, ** p<0.05, * p<0.1; standard errors in parenthesis.

Y = Teacher grade in Mathematics (from 1 to 10)	**Null model**	(S.E.)	**Model 1**	(S.E.)	**Model 2**	(S.E.)	**Model 3**	(S.E.)
Student characteristics (I level)								
Female (ref. Male)			0.510***	(0.014)	0.483***	(0.014)	0.481***	(0.015)
INVALSI test score in Numeracy (standardized)			0.901***	(0.009)	0.934***	(0.009)	0.934***	(0.009)
Quarter of birth (ref. 1st quarter)								
2nd quarter			-0.042**	(0.017)	-0.041**	(0.017)	-0.041**	(0.017)
3rd quarter			-0.056***	(0.017)	-0.057***	(0.017)	-0.057***	(0.017)
4th quarter			-0.052***	(0.017)	-0.054***	(0.017)	-0.053***	(0.017)
Migration background (ref. Native)								
2nd generation immigrant			-0.151***	(0.028)	-0.146***	(0.028)	-0.144***	(0.028)
1st generation immigrant			-0.036	(0.035)	-0.033	(0.035)	-0.032	(0.035)
Geographical area (ref. North-West)								
North-East			0.011	(0.049)	0.001	(0.044)	-0.0004	(0.044)
Center			0.107**	(0.050)	0.108**	(0.045)	0.109**	(0.045)
South			-0.033	(0.046)	-0.032	(0.042)	-0.031	(0.042)
Isles			0.140**	(0.068)	0.090	(0.062)	0.090	(0.062)
ESCS (ref. Lower)								
Medium-low			0.058***	(0.017)	0.077***	(0.017)	0.077***	(0.016)
Medium-high			0.062***	(0.018)	0.090***	(0.018)	0.092***	(0.018)
Higher			0.068***	(0.019)	0.110***	(0.019)	0.112***	(0.019)
Regularity (ref. Regular student)								
Early starter			-0.019	(0.051)	-0.008	(0.051)	-0.007	(0.051)
Late starter			-0.323***	(0.032)	-0.352***	(0.032)	-0.351***	(0.032)
Kindergarten yes (Ref. No)			0.011	(0.031)	0.015	(0.031)	0.014	(0.031)
Academic Year 2016/2017 (Ref. 2015/2016)			0.001	(0.032)	0.021	(0.029)	0.021	(0.029)
Mathematics teacher characteristics (II level)								
Female (ref. Male)					-0.136***	(0.028)	-0.136***	(0.028)
Age					-0.005***	(0.002)	-0.005***	(0.002)
Within-school seniority (ref. 1 year or less)								
2-3 years					-0.052	(0.048)	-0.054	(0.048)
4-5 years					-0.051	(0.058)	-0.053	(0.058)
More than 5 years					-0.085**	(0.043)	-0.088**	(0.043)
Permanent contract (ref. Fixed-term contract)					0.031	(0.051)	0.030	(0.051)
Classroom composition (II level)								
Class size					-0.022***	(0.003)	-0.022***	(0.003)
Percentage of females					0.005***	(0.001)	0.005***	(0.001)
Percentage of students with high ESCS					-0.002***	(0.0004)	-0.002***	(0.0004)
School type (III level)								
School track (ref. Vocational school)								
Technical school					-0.367***	(0.039)	-0.367***	(0.039)
Lyceum					-0.401***	(0.045)	-0.403***	(0.045)
Intercept	6.051***	(0.015)	5.877***	(0.052)	6.696***	(0.111)	6.693***	(0.111)
Observations	38,957		38,957		38,957		38,957	
Number of groups (class)	2,851		2,851		2,851		2,851	
Number of groups (school)	1,574		1,574		1,574		1,574	
Student Variance (level 1)	1.685		1.258		1.255		1.241	
Classroom Variance (level 2)	0.181		0.248		0.235		0.241	
School Variance (level 3)	0.171		0.190		0.121		0.121	
Variance slope (gender)							0.074	
Covariance intercept-slope							-0.020	
BIC	134442.2		124348		123983.4		123948.9	
AIC	134433.7		124185.2		123726.3		123691.8	
Log Likelihood	-67215.83		-62073.58		-61833.15		-61815.89	
Cohen's D (student gender)			0.358		0.339		0.338	

*** p<0.01, ** p<0.05, * p<0.1

Table A3.2: Multilevel regression model on teacher grade in Language; model 1: random intercept model with student-level explanatory variables; model 2: random intercept model adding classroom-level and school-level explanatory variables; model 3: random gender coefficient model with student-level, classroom-level and school-level explanatory variables. *** p<0.01, ** p<0.05, * p<0.1; standard error in parenthesis.

Y = Teacher grade in Language (from 1 to 10)	**Null Model**	(S.E.)	**Model 1**	(S.E.)	**Model 2**	(S.E.)	**Model 3**	(S.E.)
Student characteristics (I level)								
Female (ref. Male)			0.294***	(0.009)	0.301***	(0.010)	0.299***	(0.011)
INVALSI test score in Literacy (standardized)			0.537***	(0.006)	0.543***	(0.006)	0.543***	(0.006)
Quarter of birth (ref. 1st quarter)								
2nd quarter			-0.008	(0.012)	-0.008	(0.012)	-0.007	(0.012)
3rd quarter			-0.034***	(0.012)	-0.034***	(0.012)	-0.034***	(0.012)
4th quarter			-0.065***	(0.012)	-0.065***	(0.012)	-0.065***	(0.012)
Migration background (ref. Native)								
2nd generation immigrant			-0.150***	(0.020)	-0.149***	(0.020)	-0.148***	(0.020)
1st generation immigrant			-0.051**	(0.024)	-0.050**	(0.024)	-0.050**	(0.024)
Geographical area (ref. North-West)								
North-East			0.089***	(0.034)	0.081**	(0.034)	0.082**	(0.034)
Center			0.019	(0.035)	0.015	(0.035)	0.015	(0.035)
South			-0.123***	(0.032)	-0.116***	(0.032)	-0.115***	(0.032)
Isles			-0.072	(0.047)	-0.072	(0.047)	-0.072	(0.047)
ESCS (ref. Lower)								
Medium-low			0.073***	(0.012)	0.076***	(0.012)	0.076***	(0.011)
Medium-high			0.098***	(0.012)	0.102***	(0.012)	0.102***	(0.012)
Higher			0.134***	(0.013)	0.140***	(0.014)	0.140***	(0.014)
Regularity (ref. Regular student)								
Early starter			-0.045	(0.035)	-0.045	(0.035)	-0.048	(0.035)
Late starter			-0.209***	(0.022)	-0.217***	(0.022)	-0.218***	(0.022)
Kindergarten yes (Ref. No)			0.035	(0.022)	0.036*	(0.022)	0.036*	(0.022)
Academic year 2016/2017 (Ref. 2015/2016)			0.035	(0.022)	0.036	(0.022)	0.036	(0.022)
Language teacher characteristics (II level)								
Female (ref. Male)					-0.132***	(0.027)	-0.132***	(0.027)
Age					-0.002	(0.001)	-0.002	(0.001)
Within-school seniority (ref. 1 year or less)								
2-3 years					-0.025	(0.036)	-0.025	(0.036)
4-5 years					0.031	(0.043)	0.030	(0.043)
More than 5 years					-0.002	(0.034)	-0.003	(0.034)
Permanent contract (ref. Fixed-term contract)					-0.056	(0.038)	-0.055	(0.038)
Classroom composition (II level)								
Class size					-0.006***	(0.002)	-0.006***	(0.002)
Percentage of females					-0.001**	(0.0003)	-0.001**	(0.0003)
Percentage of students with high ESCS					0.0002	(0.0003)	0.0003	(0.0003)
School type (III level)								
School track (ref. Vocational school)								
Technical school					-0.087***	(0.031)	-0.086***	(0.031)
Lyceum					-0.069*	(0.035)	-0.066*	(0.035)
Intercept	6.325***	(0.013)	6.171***	(0.036)	6.601***	(0.083)	6.598***	(0.084)
Observations	38,957		38,957		38,957		38,957	
Number of groups (class)	2,851		2,851		2,851		2,851	
Number of groups (school)	1,574		1,574		1,574		1,574	
Student Variance (level 1)	0.778		0.606		0.606		0.594	
Classroom Variance (level 2)	0.182		0.187		0.185		0.199	
School Variance (level 3)	0.129		0.057		0.054		0.054	
Variance slope (gender)							0.070	
Covariance intercept-slope							-0.019	
BIC	105774		96378.77		96429.49		96317.81	
AIC	105765.4		96215.94		96172.38		96060.71	
Log Likelihood	-52881.72		-48088.97		-48056.19		-48000.35	
Cohen's D (student gender)			0.283		0.290		0.288	

*** p<0.01, ** p<0.05, * p<0.1

Figure A3.1: Random slopes calculated at the classroom level for Language and Mathematics; the models do not include teachers, classroom and school characteristics.

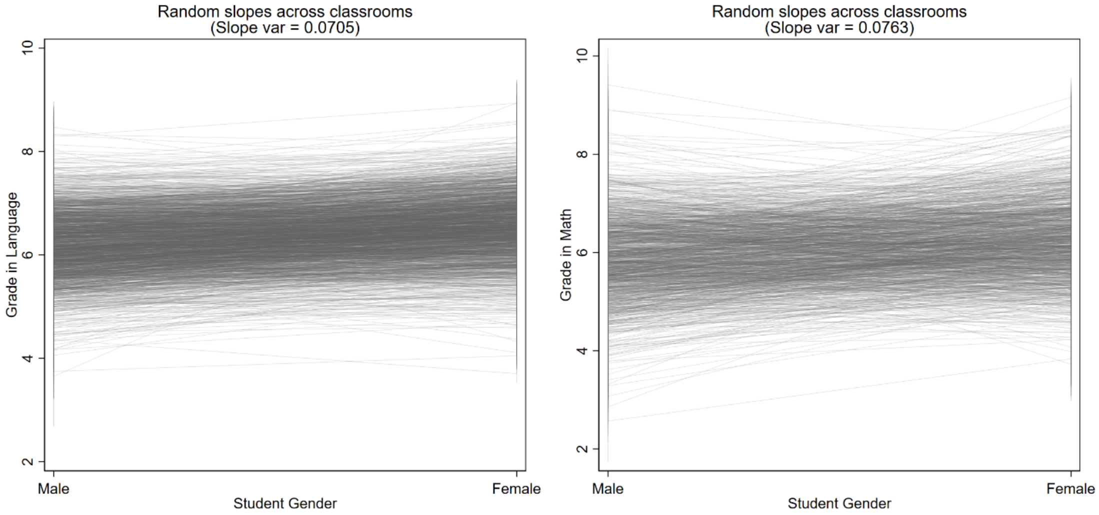


Table A3.3: Cross-level interaction terms on teacher grade in Mathematics (student gender and teacher characteristics, classroom composition and school type); coefficients for student-level control variables and academic year are omitted. *** p<0.01, ** p<0.05, * p<0.1; standard error in parenthesis.

Y = Teacher grade in Maths (from 1 to 10)	**Model 4**	(S.E.)	**Model 5**	(S.E.)	**Model 6**	(S.E.)	**Model 7**	(S.E.)	**Model 8**	(S.E.)	**Model 9**	(S.E.)	**Model 10**	(S.E.)	**Model 11**	(S.E.)
Student characteristics (I level)																
Female (ref. Male)	0.480***	(0.027)	0.535***	(0.093)	0.507***	(0.036)	0.516***	(0.048)	0.375***	(0.050)	0.425***	(0.038)	0.489***	(0.018)	0.361***	(0.035)
INVALSI test score in Numeracy (standardized)	0.934***	(0.009)	0.934***	(0.009)	0.934***	(0.009)	0.934***	(0.009)	0.934***	(0.009)	0.934***	(0.009)	0.934***	(0.009)	0.934***	(0.009)
Quarter of birth (ref. 1st quarter)																
2nd quarter	-0.041**	(0.017)	-0.041**	(0.017)	-0.041**	(0.017)	-0.041**	(0.017)	-0.042**	(0.017)	-0.041**	(0.017)	-0.041**	(0.017)	-0.042**	(0.017)
3rd quarter	-0.057***	(0.017)	-0.057***	(0.017)	-0.056***	(0.017)	-0.057***	(0.017)	-0.057***	(0.017)	-0.057***	(0.017)	-0.056***	(0.017)	-0.057***	(0.017)
4th quarter	-0.053***	(0.017)	-0.053***	(0.017)	-0.053***	(0.017)	-0.053***	(0.017)	-0.054***	(0.017)	-0.053***	(0.017)	-0.053***	(0.017)	-0.053***	(0.017)
Migration background (ref. Native)																
2nd generation immigrant	-0.144***	(0.028)	-0.144***	(0.028)	-0.144***	(0.028)	-0.144***	(0.028)	-0.144***	(0.028)	-0.144***	(0.028)	-0.144***	(0.028)	-0.144***	(0.028)
1st generation immigrant	-0.032	(0.035)	-0.032	(0.035)	-0.032	(0.035)	-0.032	(0.035)	-0.031	(0.035)	-0.032	(0.035)	-0.033	(0.035)	-0.031	(0.035)
Geographical area (ref. North-West)																
North-East	-0.0004	(0.044)	-0.0004	(0.044)	-0.0004	(0.044)	-0.0002	(0.044)	-0.001	(0.044)	-0.0003	(0.044)	-0.0001	(0.044)	-0.004	(0.044)
Center	0.109**	(0.045)	0.109**	(0.045)	0.108**	(0.045)	0.109**	(0.045)	0.109**	(0.045)	0.110**	(0.045)	0.109**	(0.045)	0.108**	(0.045)
South	-0.031	(0.042)	-0.031	(0.042)	-0.032	(0.042)	-0.031	(0.042)	-0.030	(0.042)	-0.031	(0.042)	-0.031	(0.042)	-0.031	(0.042)
Isles	0.090	(0.062)	0.090	(0.062)	0.090	(0.062)	0.090	(0.062)	0.090	(0.062)	0.090	(0.062)	0.090	(0.062)	0.088	(0.061)
ESCS (ref. Lower)																
Medium-low	0.077***	(0.016)	0.077***	(0.016)	0.077***	(0.016)	0.077***	(0.016)	0.077***	(0.016)	0.077***	(0.016)	0.077***	(0.017)	0.076***	(0.017)
Medium-high	0.092***	(0.018)	0.092***	(0.018)	0.092***	(0.018)	0.092***	(0.018)	0.091***	(0.018)	0.092***	(0.018)	0.092***	(0.018)	0.091***	(0.018)
Higher	0.112***	(0.019)	0.112***	(0.019)	0.112***	(0.019)	0.112***	(0.019)	0.111***	(0.019)	0.112***	(0.019)	0.112***	(0.019)	0.111***	(0.019)
Regularity (ref. Regular student)																
Early starter	-0.007	(0.051)	-0.008	(0.051)	-0.007	(0.051)	-0.008	(0.051)	-0.008	(0.051)	-0.007	(0.051)	-0.008	(0.051)	-0.006	(0.051)
Late starter	-0.351***	(0.032)	-0.351***	(0.032)	-0.351***	(0.032)	-0.351***	(0.032)	-0.352***	(0.032)	-0.350***	(0.032)	-0.350***	(0.032)	-0.354***	(0.032)
Kindergarten yes (Ref. No)	0.014	(0.031)	0.013	(0.031)	0.013	(0.031)	0.013	(0.031)	0.014	(0.031)	0.013	(0.031)	0.014	(0.031)	0.013	(0.031)
Academic year 2016/2017 (Ref. 2015/2016)	0.021	(0.029)	0.021	(0.029)	0.021	(0.029)	0.021	(0.029)	0.021	(0.029)	0.020	(0.029)	0.021	(0.029)	0.021	(0.029)
Maths teacher characteristics (II level)																
Female (ref. Male)	-0.137***	(0.031)	-0.136***	(0.028)	-0.136***	(0.028)	-0.136***	(0.028)	-0.136***	(0.028)	-0.136***	(0.028)	-0.136***	(0.028)	-0.134***	(0.028)
Age	-0.005***	(0.002)	-0.005**	(0.002)	-0.005***	(0.002)	-0.005***	(0.002)	-0.005***	(0.002)	-0.005***	(0.002)	-0.005***	(0.002)	-0.005***	(0.002)
Within-school seniority (ref. 1 year or less)																
2-3 years	-0.054	(0.048)	-0.053	(0.048)	-0.012	(0.055)	-0.054	(0.048)	-0.053	(0.048)	-0.054	(0.048)	-0.054	(0.048)	-0.052	(0.048)
4-5 years	-0.053	(0.058)	-0.053	(0.058)	-0.044	(0.066)	-0.053	(0.058)	-0.052	(0.058)	-0.053	(0.058)	-0.053	(0.058)	-0.049	(0.058)
More than 5 years	-0.088**	(0.043)	-0.088**	(0.043)	-0.077	(0.047)	-0.088**	(0.043)	-0.087**	(0.043)	-0.086**	(0.043)	-0.088**	(0.043)	-0.087**	(0.043)
Permanent contract (ref. Fixed-term contract)	0.030	(0.051)	0.029	(0.051)	0.030	(0.051)	0.048	(0.057)	0.030	(0.051)	0.029	(0.051)	0.030	(0.051)	0.032	(0.051)
Classroom composition (II level)																
Class size	-0.022***	(0.003)	-0.022***	(0.003)	-0.022***	(0.003)	-0.022***	(0.003)	-0.025***	(0.003)	-0.021***	(0.003)	-0.022***	(0.003)	-0.022***	(0.003)
Percentage of females	0.005***	(0.001)	0.005***	(0.001)	0.005***	(0.001)	0.005***	(0.001)	0.005***	(0.001)	0.004***	(0.001)	0.005***	(0.001)	0.005***	(0.001)
Percentage of students with high ESCS	-0.002***	(0.0004)	-0.002***	(0.0004)	-0.002***	(0.0004)	-0.002***	(0.0004)	-0.002***	(0.0004)	-0.002***	(0.0004)	-0.002***	(0.0004)	-0.002***	(0.0004)
School type (III level)																
School track (ref. Vocational school)																

Technical school	-0.367***	(0.039)	-0.368***	(0.039)	-0.368***	(0.039)	-0.368***	(0.039)	-0.367***	(0.039)	-0.370***	(0.039)	-0.367***	(0.039)	-0.430***	(0.043)
Lyceum	-0.403***	(0.045)	-0.404***	(0.045)	-0.403***	(0.045)	-0.404***	(0.045)	-0.405***	(0.045)	-0.405***	(0.045)	-0.403***	(0.045)	-0.485***	(0.050)
Cross-level interactions																
Female student#Female teacher	0.002	(0.031)														
Female student#Teacher age			-0.001	(0.002)												
Female student#2-3 years of seniority					-0.087	(0.056)										
Female student#4-5 years of seniority					-0.018	(0.065)										
Female student#more than 5 y of seniority					-0.023	(0.039)										
Female student#Permanent contract							-0.038	(0.050)								
Female#Classroom size									0.006**	(0.003)						
Female# % of females											0.001	(0.001)				
Female# % of students with high ESCS													-0.0003	(0.0003)		
Female#Techical schools															0.139***	(0.043)
Female#Lyceums															0.150***	(0.040)
Intercept	6.693***	(0.112)	6.667***	(0.120)	6.680***	(0.112)	6.676***	(0.114)	6.734***	(0.113)	6.704***	(0.112)	6.691***	(0.112)	6.741***	(0.112)
Observations	38,957		38,957		38,957		38,957		38,957		38,957		38,957		38,957	
Number of groups (class)	2.851		2.851		2.851		2.851		2.851		2.851		2.851		2.851	
Number of groups (school)	1,574		1,574		1,574		1,574		1,574		1,574		1,574		1,574	
Variance slope (gender)	0.074		0.073		0.073		0.073		0.074		0.074		0.073		0.072	
Student Variance (level 1)	1.241		1.241		1.241		1.241		1.241		1.241		1.241		1.242	
Classroom Variance (level 2)	0.241		0.241		0.240		0.241		0.240		0.241		0.241		0.239	
School Variance (level 3)	0.121		0.121		0.121		0.121		0.121		0.121		0.122		0.120	
BIC	123959.5		123959.1		123978		123958.9		123954.4		123956.8		123958.7		123955.2	
AIC	123693.8		123693.4		123695.2		123693.2		123688.7		123691.2		123693		123681	
Log Likelihood	-61815.89		-61815.72		-61814.6		-61815.6		-61813.36		-61814.58		-61815.52		-61808.5	
Cohen's D (student gender)	0.338		0.338		0.338		0.338		0.336		0.336		0.339		0.338	
Cohen's D (student gender # 0/lower)	0.337		0.356		0.356		0.362		0.285		0.306		0.341		0.253	
Cohen's D (student gender # 1/higher)	0.338		0.323		0.340		0.335		0.394		0.367		0.325		0.351	
Cohen's D (student gender # 2)															0.358	

*** p<0.01, ** p<0.05, * p<0.1

Table A3.4: Cross-level interaction terms on teacher grade in Language (student gender and teacher characteristics, classroom composition and school type); coefficients for student-level control variables and academic year are omitted. *** p<0.01, ** p<0.05, * p<0.1; standard error in parenthesis.

Y = Teacher grade in Language (from 1 to 10)	**Model 4**	(S.E.)	**Model 5**	(S.E.)	**Model 6**	(S.E.)	**Model 7**	(S.E.)	**Model 8**	(S.E.)	**Model 9**	(S.E.)	**Model 10**	(S.E.)	**Model 11**	(S.E.)
Student characteristics (I level)																
Female (ref. Male)	0.327***	(0.028)	0.227***	(0.064)	0.268***	(0.026)	0.246***	(0.034)	0.275***	(0.037)	0.306***	(0.028)	0.301***	(0.013)	0.259***	(0.026)
INVALSI test score in Literacy (standardized)	0.543***	(0.006)	0.543***	(0.006)	0.543***	(0.006)	0.543***	(0.006)	0.543***	(0.006)	0.543***	(0.006)	0.543***	(0.006)	0.543***	(0.006)
Quarter of birth (ref. 1st quarter)																
2nd quarter	-0.007	(0.012)	-0.007	(0.012)	-0.007	(0.012)	-0.007	(0.012)	-0.007	(0.012)	-0.007	(0.012)	-0.007	(0.012)	-0.007	(0.012)
3rd quarter	-0.034***	(0.012)	-0.034***	(0.012)	-0.034***	(0.012)	-0.034***	(0.012)	-0.034***	(0.012)	-0.034***	(0.012)	-0.034***	(0.012)	-0.034***	(0.012)
4th quarter	-0.065***	(0.012)	-0.065***	(0.012)	-0.065***	(0.012)	-0.065***	(0.012)	-0.065***	(0.012)	-0.065***	(0.012)	-0.065***	(0.012)	-0.065***	(0.012)
Migration background (ref. Native)																
2nd generation immigrant	-0.148***	(0.020)	-0.148***	(0.020)	-0.148***	(0.020)	-0.148***	(0.020)	-0.148***	(0.020)	-0.148***	(0.020)	-0.148***	(0.020)	-0.148***	(0.020)
1st generation immigrant	-0.050**	(0.024)	-0.050**	(0.024)	-0.050**	(0.024)	-0.050**	(0.024)	-0.050**	(0.024)	-0.050**	(0.024)	-0.050**	(0.024)	-0.050**	(0.024)
Geographical area (ref. North-West)																
North-East	0.082**	(0.034)	0.082**	(0.034)	0.082**	(0.034)	0.082**	(0.034)	0.082**	(0.034)	0.082**	(0.034)	0.082**	(0.034)	0.081**	(0.034)
Center	0.016	(0.035)	0.016	(0.035)	0.015	(0.035)	0.015	(0.035)	0.015	(0.035)	0.015	(0.035)	0.015	(0.035)	0.015	(0.035)
South	-0.115***	(0.032)	-0.115***	(0.032)	-0.116***	(0.032)	-0.115***	(0.032)	-0.115***	(0.032)	-0.115***	(0.032)	-0.115***	(0.032)	-0.115***	(0.032)
Isles	-0.072	(0.047)	-0.072	(0.047)	-0.072	(0.047)	-0.072	(0.047)	-0.072	(0.047)	-0.072	(0.047)	-0.072	(0.047)	-0.073	(0.047)
ESCS (ref. Lower)																
Medium-low	0.076***	(0.011)	0.076***	(0.011)	0.076***	(0.011)	0.076***	(0.011)	0.076***	(0.011)	0.076***	(0.011)	0.076***	(0.011)	0.075***	(0.011)
Medium-high	0.102***	(0.012)	0.102***	(0.012)	0.102***	(0.012)	0.102***	(0.012)	0.102***	(0.012)	0.102***	(0.012)	0.102***	(0.012)	0.102***	(0.012)
Higher	0.140***	(0.014)	0.140***	(0.014)	0.140***	(0.014)	0.140***	(0.014)	0.140***	(0.014)	0.140***	(0.014)	0.140***	(0.014)	0.140***	(0.014)
Regularity (ref. Regular student)																
Early starter	-0.048	(0.035)	-0.048	(0.035)	-0.048	(0.035)	-0.048	(0.035)	-0.048	(0.035)	-0.048	(0.035)	-0.048	(0.035)	-0.048	(0.035)
Late starter	-0.218***	(0.022)	-0.218***	(0.022)	-0.218***	(0.022)	-0.218***	(0.022)	-0.218***	(0.022)	-0.218***	(0.022)	-0.218***	(0.022)	-0.219***	(0.022)
Kindergarten yes (Ref. No)	0.036*	(0.022)	0.037*	(0.022)	0.037*	(0.022)	0.037*	(0.022)	0.036*	(0.022)	0.037*	(0.022)	0.036*	(0.022)	0.036*	(0.022)
Academic year 2016/2017 (Ref. 2015/2016)	0.036	(0.022)	0.036	(0.022)	0.036	(0.022)	0.036	(0.022)	0.036	(0.022)	0.036	(0.022)	0.036	(0.022)	0.036	(0.022)
Language teacher characteristics (II level)																
Female (ref. Male)	-0.117***	(0.031)	-0.132***	(0.027)	-0.132***	(0.027)	-0.132***	(0.027)	-0.132***	(0.027)	-0.132***	(0.027)	-0.132***	(0.027)	-0.131***	(0.027)
Age	-0.002	(0.001)	-0.002	(0.002)	-0.002	(0.001)	-0.002	(0.001)	-0.002	(0.001)	-0.002	(0.001)	-0.002	(0.001)	-0.002	(0.001)
Within-school seniority (ref. 1 year or less)																
2-3 years	-0.025	(0.036)	-0.024	(0.036)	-0.053	(0.040)	-0.025	(0.036)	-0.024	(0.036)	-0.025	(0.036)	-0.025	(0.036)	-0.024	(0.036)
4-5 years	0.030	(0.043)	0.030	(0.043)	0.0004	(0.047)	0.029	(0.043)	0.030	(0.043)	0.030	(0.043)	0.030	(0.043)	0.030	(0.043)
More than 5 years	-0.003	(0.034)	-0.003	(0.034)	-0.017	(0.037)	-0.004	(0.034)	-0.003	(0.034)	-0.003	(0.034)	-0.003	(0.034)	-0.003	(0.034)
Permanent contract (ref. Fixed-term contract)	-0.055	(0.038)	-0.056	(0.038)	-0.056	(0.038)	-0.084**	(0.042)	-0.056	(0.038)	-0.055	(0.038)	-0.055	(0.038)	-0.056	(0.038)
Classroom composition (II level)																
Class size	-0.006***	(0.002)	-0.006***	(0.002)	-0.006***	(0.002)	-0.006***	(0.002)	-0.007***	(0.003)	-0.006***	(0.002)	-0.006***	(0.002)	-0.006***	(0.002)
Percentage of females	-0.001**	(0.0004)	-0.001**	(0.0004)	-0.001**	(0.0004)	-0.001**	(0.0004)	-0.001**	(0.0004)	-0.001**	(0.0005)	-0.001**	(0.0004)	-0.001**	(0.0004)
Percentage of students with high ESCS	0.0003	(0.0003)	0.0003	(0.0003)	0.0003	(0.0003)	0.0003	(0.0003)	0.0003	(0.0003)	0.0003	(0.0003)	0.0003	(0.0003)	0.0003	(0.0003)
School type (III level)																
School track (ref. Vocational school)																
Technical school	-0.086***	(0.031)	-0.086***	(0.031)	-0.085***	(0.031)	-0.085***	(0.031)	-0.086***	(0.031)	-0.085***	(0.031)	-0.086***	(0.031)	-0.112***	(0.034)

Lyceum	-0.066*	(0.035)	-0.066*	(0.035)	-0.066*	(0.035)	-0.066*	(0.035)	-0.067*	(0.035)	-0.066*	(0.035)	-0.066*	(0.035)	-0.090**	(0.039)
Cross-level interactions																
Female student#Female teacher	-0.033	(0.030)														
Female student#Teacher age			0.001	(0.001)												
Female student#2-3 years of seniority					0.062	(0.039)										
Female student#4-5 years of seniority					0.062	(0.044)										
Female student#more than 5 y of seniority					0.029	(0.030)										
Female student#Permanent contract							0.059*	(0.035)								
Female#Classroom size									0.001	(0.002)						
Female# % of females											-0.0001	(0.0005)				
Female# % of students with high ESCS													-0.0006	(0.0003)		
Female#Techical schools															0.059*	(0.032)
Female#Lyceums															0.044	(0.030)
Intercept	6.586***	(0.084)	6.631***	(0.089)	6.611***	(0.084)	6.621***	(0.085)	6.607***	(0.085)	6.596***	(0.084)	6.597***	(0.084)	6.616***	(0.084)
Observations	38,957		38,957		38,957		38,957		38,957		38,957		38,957		38,957	
Number of groups (class)	2,851		2,851		2,851		2,851		2,851		2,851		2,851		2,851	
Number of groups (school)	1,574		1,574		1,574		1,574		1,574		1,574		1,574		1,574	
Variance slope (gender)	0.070		0.070		0.070		0.070		0.070		0.070		0.070		0.070	
Student Variance (level 1)	0.594		0.594		0.594		0.594		0.594		0.594		0.594		0.594	
Classroom Variance (level 2)	0.199		0.199		0.199		0.199		0.199		0.199		0.199		0.199	
School Variance (level 3)	0.054		0.054		0.054		0.054		0.054		0.054		0.054		0.054	
BIC	96327.19		96327.06		96346.21		96325.56		96327.91		96328.31		96328.33		96335.54	
AIC	96061.51		96061.39		96063.39		96059.88		96062.24		96062.64		96062.65		96061.29	
Log Likelihood	-47999.76		-47999.69		-47998.7		-47998.94		-48000.12		-48000.32		-48000.32		-47998.65	
Cohen’s D (student gender)	0.288		0.288		0.288		0.288		0.288		0.288		0.288		0.289	
Cohen’s D (student gender # 0/lower)	0.314		0.256		0.258		0.237		0.279		0.293		0.289		0.249	
Cohen’s D (student gender # 1/higher)	0.283		0.310		0.286		0.294		0.306		0.283		0.284		0.306	
Cohen’s D (student gender # 3)															0.291	

*** p<0.01, ** p<0.05, * p<0.1

Table A3.5: Comparison across different model specification. MLM (multilevel linear model) with random slope on teacher grade in Mathematics (selected model); FEM (Linear regression model with fixed-effects at the classroom level); MLM+IPW (multilevel linear regression model with weighted coefficients); MLM+IV (multilevel linear regression model with previous INVALSI test score as instrumental variable). Coefficients for academic year are omitted. *** p<0.01, ** p<0.05, * p<0.1; standard error in parenthesis.

Y = Teacher grade in Mathematics (from 1 to 10)	**MLM**	(S.E.)	**FEM**	(S.E.)	**MLM + IPW**	(S.E.)	**MLM + IV**	(S.E.)
Student characteristics (I level)								
Female (ref. Male)	0.481***	(0.015)	0.497***	(0.014)	0.483***	(0.016)	0.526***	(0.021)
INVALSI test score in Numeracy (standardized)	0.934***	(0.009)	1.017***	(0.010)	0.926***	(0.014)	1.363***	(0.023)
Quarter of birth (ref. 1st quarter)								
2nd quarter	-0.041**	(0.017)	-0.043**	(0.017)	-0.043**	(0.017)	-0.017	(0.023)
3rd quarter	-0.057***	(0.017)	-0.056***	(0.017)	-0.060***	(0.017)	-0.039*	(0.022)
4th quarter	-0.053***	(0.017)	-0.054***	(0.017)	-0.056***	(0.017)	-0.041*	(0.023)
Migration background (ref. Native)								
2nd generation immigrant	-0.144***	(0.028)	-0.145***	(0.029)	-0.149***	(0.030)	-0.158***	(0.038)
1st generation immigrant	-0.032	(0.035)	-0.034	(0.035)	-0.033	(0.040)	-0.047	(0.046)
Geographical area (ref. North-West)								
North-East	-0.0004	(0.044)	-	-	0.003	(0.042)	0.003	(0.047)
Center	0.109**	(0.045)	-	-	0.110**	(0.043)	0.216***	(0.049)
South	-0.031	(0.042)	-	-	-0.028	(0.043)	0.170***	(0.046)
Isles	0.090	(0.062)	-	-	0.090	(0.065)	0.351***	(0.071)
ESCS (ref. Lower)								
Medium-low	0.077***	(0.016)	0.076***	(0.017)	0.079***	(0.017)	0.048**	(0.022)
Medium-high	0.092***	(0.018)	0.087***	(0.018)	0.091***	(0.018)	0.053**	(0.024)
Higher	0.112***	(0.019)	0.107***	(0.020)	0.114***	(0.020)	0.033	(0.027)
Regularity (ref. Regular student)								
Early starter	-0.007	(0.051)	-0.018	(0.051)	-0.004	(0.050)	-0.104	(0.068)
Late starter	-0.351***	(0.032)	-0.353***	(0.032)	-0.356***	(0.035)	-0.271***	(0.043)
Kindergarten yes (Ref. No)	0.014	(0.031)	0.005	(0.031)	0.014	(0.032)	-0.068*	(0.040)
Academic year 2016/2017 (Ref. 2015/2016)	0.021	(0.029)	-	-	0.019	(0.028)	0.038	(0.042)
Mathematics teacher characteristics (II level)								
Female (ref. Male)	-0.136***	(0.028)	-	-	-0.134***	(0.029)	-0.175***	(0.031)
Age	-0.005***	(0.002)	-	-	-0.005***	(0.002)	-0.001	(0.002)
Seniority (ref. 1 year or less)								
2-3 years	-0.054	(0.048)	-	-	-0.048	(0.048)	-0.021	(0.055)
4-5 years	-0.053	(0.058)	-	-	-0.051	(0.058)	-0.127**	(0.064)
More than 5 years	-0.088**	(0.043)	-	-	-0.084*	(0.045)	-0.206***	(0.049)
Permanent contract (ref. Fixed-term contract)	0.030	(0.051)	-	-	0.031	(0.052)	0.025	(0.059)
Classroom composition (II level)								
Class size	-0.022***	(0.003)	-	-	-0.022***	(0.003)	-0.035***	(0.003)
Percentage of females	0.005***	(0.001)	-	-	0.005***	(0.001)	0.009***	(0.001)
Percentage of students with high ESCS	-0.002***	(0.0004)	-	-	-0.002***	(0.0004)	-0.002***	(0.0004)
School type (III level)								
School track (ref. Vocational school)								
Technical school	-0.367***	(0.039)	-	-	-0.369***	(0.040)	-0.464***	(0.044)
Lyceum	-0.403***	(0.045)	-	-	-0.402***	(0.047)	-0.662***	(0.051)
				-				
Intercept	6.693***	(0.111)	5.829***	(0.035)	6.685***	(0.117)	6.702***	(0.126)
Observations	38,957		38,957		38,957		23,581	
Number of groups (class)	2,851		-		2,851		2,309	
Number of groups (school)	1,574		-		1,574		1,349	
Intercept Variance (Between) (level 2)	0.074		-		0.033		0.081	
Residual Variance (Within) (level 1)	0.121		-		0.120		0.073	
Variance slope (gender)	1.241		-		1.259		1.383	
Covariance intercept-slope	-0.020		-		-0.006		-0.025	
BIC	123948.9		116490.4		112640		77177.33	
AIC	123691.8		116370.5		112383.2		76935.28	
Log Likelihood	-61815.89		-58171.23		-56161.58		-38437.64	

*** p<0.01, ** p<0.05, * p<0.1

CHAPTER 4

STUDENTS' PROFILES AND SOCIAL STRATIFICATION OF ADOLESCENTS' SOCIOEMOTIONAL SKILLS: A COMPREHENSIVE UNDERSTANTING OF GRADE DETERMINANTS

Abstract

Teacher grades are a multidimensional assessment of students' academic ability, blending a number of different factors. Among these, students' capacity to engage in the social process of schooling according to their socioemotional and noncognitive skills has been often neglected. The aim of this chapter is filling the gap in the sociological literature about grade determinants, by considering the interplay of students' socioemotional skills among each other, and their social stratification. I rely on a novel dataset in which Italian INVALSI-SNV data are merged with PISA OECD 2018 data, containing information on 6,504 15-years-old Italian students. Results show that students' profiles significantly predict teacher grades also when controlling for students' ascriptive characteristics and academic competences. The relationship between students' characteristics and teacher grade in Language is similar according to students' profiles, while in Mathematics it varies.

Keywords: socioemotional skills; non-cognitive skills; school attitude; teacher grades; latent profile analysis.

Introduction

Teacher grade is the most common measure of educational outcomes. For many decades, the issue of what teacher grades should measure, and of what they do measure, has been a core discussion among scholars as well as policy makers, principals and educational experts. The central question binding all previous studies on the topic is "what do grades mean?" (Brookhart et al. 2016), and this reflects the urgency to understand whether this measure is correctly intended and implemented in educational contexts. The investigation of grades is fundamental, as teacher assigned grades have a central role in students' educational journey. Indeed, grades are signals used by parents, schools and teachers to indicate students' academic ability and their possible educational future (Pattison et al. 2013). Grades are important predictor of a variety of educational outcomes such as school dropout (Bowers et al. 2013) and college success (Thorsen & Cliffordson 2012) as well as labor market outcomes, such as future earnings and occupational choices (Lavy & Sand 2015; Bonner & Chen 2019). Furthermore, grades are cues used by students themselves, and especially during critical ages such adolescence, teachers' judgments of students' abilities may have a great impact on the perceptions of their own abilities and consequently on their future choices.

Despite their relevance, historically, teacher grades have been conceived as relatively unreliable measures of students' academic ability, because they explain only about 25-35% of more objective indicators such as standardized test scores (Bowers 2011). Indeed, grades are more strongly related to multiple noncognitive factors than achievement tests (Borghans et al. 2011; Farrington et al. 2012; Lechner et al. 2017; DeVries et al. 2018). Existing literature suggests that grades are a multidimensional assessment that involves both students' academic ability and competences, and students' broader capacity to engage in the social process of schooling (Klapp et al. 2009). However, teachers' evaluations of students are also imbued with social considerations related to both student and teacher sociodemographic characteristics, their interactions, and the context in which the relationship happens (Costrell 1994; Chen & Bonner 2017; Bonner & Chen 2019).

The issue of teacher assessment, or grade determinants, has been mainly studied from two perspectives: a psychological and a sociological one. Previous psychological research focused on identifying the individual dimensions predicting teacher assigned grades. The focus here is on the non-cognitive behavioral aspects such as students' attitudes and behaviours, soft skills, socioemotional skills and psychological and personality traits (see for example Egalite et al. 2016; DeVries et al. 2018; Gerbino et al. 2018). Among these, "classroom participation, effort, behavior, attendance, improvement, and turning in homework" (Bowers 2011: 1), have been found to be significantly correlated with grades rather than with standardized scores.

Scholars from sociology and economics of education focused instead on unveiling the mechanisms behind grading bias, that occurs when "a teacher gives students of different [groups] grades that systematically differ but not due to their performance" (Protivínský & Münich 2018: 141). This stream of literature mainly focused on disentangling the weight of teacher expectation bias in grading practices alongside with student ascriptive characteristics, such as gender (e.g., Lievore & Triventi 2022), ethnic background (e.g., Hinton & Higson 2017) and socioeconomic background (e.g., Bygren 2020).

Despite the relevance and the richness of previous findings about grade determinants, a bridge between the two abovementioned streams of research is still missing. Previous empirical studies failed in considering altogether: 1) the nonrandom distribution across the student population of non-cognitive and socioemotional skills, and schooling attitudes, which may be shaped by students' ascriptive characteristics (see DeVries et al. 2018); and 2) the interdependence and interplay of non-cognitive and socioemotional skills. Indeed, the focus has been mainly on finding associations between *single* students' characteristics – either non-cognitive or ascriptive - and teacher assigned grades.

Instead, it is reasonable to assume that teacher assessment is the result of the interplay of numerous factors intervening and happening altogether (Isnawati & Saukah 2017). This chapter builds on the assumption that teachers cannot sensibly distinguish a single students' characteristics or non-cognitive dimension from the interplay of the numerous and relevant cognitive, non-cognitive and ascriptive factors in the schooling setting. Indeed, imagining a classroom context in which different actions and interactions are in place simultaneously, it may be an impossible task for teachers to evaluate students according to their separate qualities.

This chpater aims at further the understanding of grade determinants, introducing a novel perspective with the purpose of enlightening the interplay of the dimensions related to student grades. The starting point is the exploitation of students' non-cognitive skills, socioemotional skills, and schooling attitudes – determinants for teacher assigned grades – to partition students in classes/profiles where within-group differences are minimized on the basis of such skills and their distribution.

Accordingly, the determined profiles are then employed to answer to the following research questions:

i) Are students' gender, migratory background and socio-economic background significant predictors for students' belonging to different profiles?

ii) Are student profiles significant predictors of teacher assigned grades? Does this relationship hold when accounting for students' ascriptive characteristics and students' academic competences?

iii) Does the relationship between students' ascriptive characteristics (gender, migratory background and socioeconomic background) and teacher assigned grades change according to students' profiles?

The empirical analysis relies on an original dataset that merges INVALSI Italian data with PISA OECD data. INVALSI data contains information on students' grades, their subject-specific academic competences and their ascriptive characteristics. PISA-OECD data contains rich information on students' non-cognitive skills, socioemotional skills and schooling attitudes. The analytical sample includes 6,504 15-years-old Italian students in 2018. Through latent profile analysis, I identify different student profiles according to their non-cognitive and socioemotional skills. I perform multivariate logistic regression analysis to describe the stratification of students across the different profiles. Finally, I rely on OLS models controlling for students' subject-specific competences to assess the relationship between student profiles and teacher assigned grades in two subject –

Language and Mathematics. This allows to capture a *total* effect[26] of students' socioemotional skills and schooling attitudes on teachers' assigned grades.

The chapter is organized as follows: the next section develops a theoretical background that describes the function of teacher grades and its dimensions. The third section illustrates the analytical strategy implemented in this chapter, including information on data, methods and variables. The fourth section presents the findings, and the last section summarizes the conclusions and proposes a discussion.

The Function of Teacher Grades

Brookhart (1991) defines grades and the practice of grading as a "hodgepodge" (1991:36) in which achievement factors, together with other factors related to students such as student effort, behaviour, attitudes, and improvement are mixed. In addition, teachers' evaluation of students come from the interplay of these numerous factors that occurs in a specific surrounding context. This context implies social considerations that shape teacher beliefs, expectations and evaluations, also through individual students and teachers' characteristics, student-teacher interactions and interactions between classmates (McMillan 2003; Randall & Engelhard 2010; Kunnath 2017).

Because of the interplay of these numerous cognitive, non-cognitive and individual characteristics, summarizing teacher grades is not straightforward. However, teacher grades may be conceptualized as a function of three different components broadly

[26] The use of the term total "effect" is not meant to imply a causal relationship. Total effect in this framework is used in contraposition to mediated effects – meaning the focus is not on the mechanisms underlying the relationship under analysis.

defined: i) students' academic/intellectual ability; ii) students' ascriptive characteristics; iii) students' non-cognitive or socioemotional skills. These components can engage in different combinations, and can interact among themselves and with both the surrounding context, and with the teacher and his/her characteristics.

The multitude of dimensions involved makes it challenging for educational researchers to understand the way in which teachers evaluate their students. This is the reason why most of the existing empirical research focuses on single student traits, trying to build up knowledge on this topic one piece at the time.

Students' Academic Ability and Characteristics

The first component in the teacher grade equation is students' cognitive factors and academic ability. One of teachers' crucial professional skills in indeed to being able to understand and to capture through grades students' ability (Ready & Wright 2011), understood as the capacity to perform cognitive tasks through a correct and appropriate processing of mental information (Carroll 1993). However, previous research shows that teachers' judgements of their students' cognitive ability is only moderately accurate (Machts et al. 2016), not all teachers are equally good in assigning grades, and some students are more likely to be assessed fairly than others (Baudson et al. 2016).

From an empirical point of view, one measure that accounts for student cognitive ability beyond grades is standardized test scores. Standardized achievement tests are tools that allow comparisons of knowledge and skills of students of the same age or grade in a defined area (Popham 1999), and they are designed to capture specific competences acquired in school (Heckman & Kautz 2014). The validity of standardized achievement tests has always been questioned in the field of educational study (for an example, see Gosling 1968) but it is also the most accurate proxy of student ability that is available in standard datasets. Achievement tests, as often blindly evaluated, are thought to be independent of both teacher expectations and teacher assessments of noncognitive traits that students display in a specific educational context (Borghans et al. 2016).

Even if a number of studies uses grades and test scores as interchangeably identifying students' competences, the two measures correlated only moderately (Willingam et al. 2002). A consistent body of research demonstrates that "noncognitive" skills are not captured by standardized tests (Jackson 2018), while grades have been shown to reflect numerous personality factors in addition to academic competence (Borghans et al. 2011; Andrei et al. 2015; Lechner et al. 2017; Gerbino et al. 2018). Therefore, even if standardized test scores may be biased to some extent (for example by test anxiety, see von der Embse et al. 2018), they are a reasonable measure of students' cognitive/academic ability, and may serve as a yardstick against which to assess differences in teacher judgments that are not explained by differences in students' academic ability.

The difference between grades and standardized test scores may be partially explained by students' ascriptive characteristics. Indeed, a recent stream of research on grade determinants focused on findings associations between students' characteristics and teacher grades comparing students with similar or identical academic competences measured through standardized test scores. The stream of research that focuses on how students' ascriptive characteristics affect the way teachers evaluate their students builds up on the teacher expectation theory, and broadly on the issue of teacher bias. Several authors investigated how students' characteristics may shape teachers' expectations (see

Jussim & Harber 2005 for an overview), and this occurs mainly through teachers' stereotypes as representations of characteristics of specific groups (Bordalo et al. 2016). Similar achievement shown by students having different background characteristics may be assessed differently, and this may depend on "suitable properties" of the group of belonging acting in the specific context (Correll & Benard 2006). The main sources shaping teachers' expectations are ethnicity, gender and socioeconomic status of students.

Concerning students' ethnic background, previous research suggests that teachers tend to evaluate the behaviour of students with a different ethnicity than their own as more disruptive, inattentive, and more likely to not be able to complete their tasks (Dee 2005). Regarding the stereotypical assessment of students according to their gender, previous studies reported perceived differences in interests, attitudes and behaviors attributed by teachers to either boys or girls (Kollmayer et al. 2018). For example, teachers usually perceive female students as more motivated, as more eager to learn (e.g., Gentrup & Rjosk 2018), as behaving better (Glock & Kleen 2017) and as having less disruptive behaviour in classroom (Terrier 2015). Finally, teachers tend to have higher expectations for children coming from higher socioeconomic background (Speybroeck et al. 2012), since they are perceived as showing more self-control and engagement in the classroom, and on average they may give an impression of brilliance successively rewarded in terms of grades by their teachers (Cole & Mendick 2006).

All the above-mentioned empirical work suggests that teacher expectations are strictly linked to what they expect to happen within the classroom context as regard to students' schooling attitudes, socioemotional skills, and non-cognitive traits. Some studies showed how non-cognitive skills, personality traits, and behaviours are not randomly distributed, but they actually differ systematically according to gender, socioeconomic status and ethnic background. Accordingly, the significant differences in a multitude of dimensions ("hodgepodge" of non-cognitive skills) between social groups tend to favour female, native students with higher SES when considering teacher assessment (Speybroeck et al. 2012; Fletcher & Wolfe 2016; DeVries et al. 2018; Nguyen et al. 2019). Indeed, these groups show specific non-cognitive factors that are positively rewarded in terms of grades and that are suited for the schooling context.

Non-Cognitive Factors, Schooling Attitudes and Socioemotional Skills

Non-cognitive skills refer to the whole set of individual behaviours, attitudes, and strategies that have been shown to be associated with a lot of indicators of individual success – along which academic ability and teacher grades. Noncognitive (or socioemotional) skills have been defined as "personality traits, goals, character, motivations, and preferences" (Kautz et al. 2014: 2) that represent individuals' patterns of behaviour. They incorporate constructs such as optimism, resilience, adaptability, and conscientiousness (Egalite et al. 2016). The empirical results concerning how students' non-cognitive and socioemotional skills affecting teacher assessed grades are mixed, and somehow fragmented.

Following Brookhart (2019), the term "factors" is used to describe the elements that teachers use as source of evidence in order to make specific judgements and to assign grades. Of course, every teacher weights differently the single "grade factors" according to his/her expectations and personality, according to the classroom experience, and also according to the grade level, where similar factors have different weights according to students' growth (Guskey & Link 2019).

Overall, research on grade determinants shows that teachers use a variety of students' behaviour, attitudes and "soft" skills when assessing their performance (Brookhart et al. 2016). For example, among students' behaviour, McMillan (2001) identifies four "academic enablers" (2001:25): effort, work habits, attention, participation. Guskey and Link (2019) include also homework competition and quality, neatness, and progress made. Among students' soft skills, previous research analyzed also the impact of grit, self-control, self-confidence (Mulchany-Dunn 2018), perseverance and passion (Egalite et al. 2016).

Some researchers focused also on how personality traits may affect teacher grades. In particular, conscientiousness and agreeableness are the ones that have a large and positive impact on academic achievement measured through grade point average (Komarraju et al. 2009; Conard 2006), while neuroticism has mainly been found to correlate negatively with teacher assessment (Laidra et al. 2007). Concerning the personality traits of extroversion and openness to experience, results are mixed and the effect on grades depend on either the subject or the grade level (Melissa et al. 2007; Furnham 2003). Among students' attitudes, social behaviours are strongly related to grades. In particular, pro-social behaviours is a strong predictor of academic grades above and beyond students' cognitive abilities (Gerbino et al. 2018), as well as peer problems (DeVries et al. 2018). Other researchers include problem behaviors in general, motivation and also life satisfaction, as factors predicting teacher grades (Enzi 2015; Angelo 2014).

Research Design

Data

In order to address the questions posed in this study, two different sources of information are needed. The first source is represented by the Italian National Institute for the Evaluation of the Education System (INVALSI) within the National Evaluation System (SNV). INVALSI-SNV performs yearly systematic assessment of students' subject-related academic competences in specific school grades[27], through tests standardized at the national level. INVALSI data includes information about students from administrative sources, and in specific grades and academic years also from students' questionnaire for the whole Italian student population. In this analysis, the focus is on information collected for students in 10th grades[28]. The INVALSI-SNV dataset serves the purposes of this study, since it contains information on both teacher assessment of students' academic ability, through teacher grades in two subjects (Language and Mathematics), and student subject-specific competences in the same subjects, measured by the INVALSI standardized test score. The two measures of teacher grades and standardized scores are extremely reliable, because instead of being self-reported by students, they are registered by administrative sources.

The second source of information derives from the OCSE PISA 2018 data. PISA (Programme for International Student Assessment) collects comprehensive information about the academic ability and knowledge of a random sample of 15-years-old students around the world. Through standardized tests, PISA measures students' ability every

[27] Grades in which INVALSI tests are administered are usually: 2nd, 5th, 8th, 10th, 15th

[28] INVALSI collects information on about 500,000 10th graders every year.

three years in more than 80 countries in different domains such as reading, mathematics, and science, with a focus on a specific subject every time PISA is administered. PISA 2018 collects also other information, including different questionnaires administered to students, as well as to school principals, teachers and parents. The PISA 2018 dataset, and in particular the students' questionnaire, contains rich information about a variety of students indicators, among which are students' psychological traits, socioemotional skills, non-cognitive skills, attitude toward school, behaviours in classroom, schedule and learning time. The PISA 2018 dataset contains information about 11,279 15-years-old Italian students. Among these, 10,680 observations are selected for having non-missing data regarding the information of interest.

In Italy, INVALSI-SNV is officially in charge of the PISA survey administration and data collection. Therefore, for the year 2018 it is possible to link information from the PISA survey with information on 10th grade students from the INVALSI survey through the SIDI code – a student unique identifier[29]. Around 2,000 students could not be linked to the PISA dataset due to missing SIDI code, and the reasons could be either a mismatch with the SIDI code or the fact that students were 15 years old in later or previous academic years: I decided not to include them because of the excessively large time span from the PISA survey and the INVALSI test. Other missing cases are due to missing information about both INVALSI test scores and teacher assessment. After the merging

[29] Through the same identifier it is also possible to link information about students' performance in their 8th grade measured by INVALSI, in order to include a measure of previous academic competences. Since this merge implies a further loss of cases, analysis including students' academic competences in 8th grade are included only in the appendix (see Robustness Checks section)

and listwise case deletion, the final sample consists of 6,504 students with all the information needed from the merge of the INVALSI dataset with the PISA dataset.

PISA collects information on 15 years-old students regardless of their academic history, or the grade in which they are enrolled at the time of the survey. At the time of PISA data collection, the vast majority of 15 years old students were enrolled in grade 10 in the academic year 2017-18 (6,808 students merged with the information needed). I further merged those students with information on both teacher grades and INVALSI test score who were 15 years old in 2018 but were enrolled in 10th grade in the academic year 2018/19 (279 students) when they were administered the INVALSI test[30]. After listwise case deletion, the final sample consists of 6,504 students with all the information needed from the merge of the INVALSI dataset with the PISA dataset.

The combined dataset using INVALSI-SNV data and OECD PISA data allows for the first time to investigate how Italian 15-years-old students' non-cognitive or socioemotional skills and attitudes toward schools, and the combination of all these skills, may affect the way in which students are evaluated by their teachers, controlling for their actual academic ability in two different subjects: Language and Mathematics.

Analytical Strategy

The analysis is organized in three main stages. The first step is the identification of students' profiles using an extensive set of indexes elaborated by PISA, as indicators of

[30] At first, this merge included also students who were in 10th grade in the academic year 2016-17 (17 students merged with the information needed). After listwise deletion, those students were excluded from the final sample because of missing information.

students' soft skills, schooling attitudes and socioemotional traits. In this stage, latent profile analysis (LPA) is performed, and different specifications are tested in order to find the best balance between model fit and parsimony.

The second step is the examination of whether these students' profiles are socially stratified by students' gender, socioeconomic background and migratory background. The goal is assessing whether students' ascriptive characteristics predict the probability of belonging to one of the determined student profiles. Multinomial logistic regression models are performed on the analytical sample of students, in which the outcome variable is belonging to student profiles, whereas the independent variables are students' gender, ESCS, and migratory background.

The third and final step is twofold: 1) assessing whether belonging to a specific students' profile is correlated to a higher or lower teacher assessment, controlling for students' current subject-specific performance and ascriptive characteristics; 2) assessing whether the relationship between students' ascriptive characteristics (gender, migratory background and socioeconomic status) and teacher assessment changes according to students' profiles.

In order to investigate how teachers assigned grades to their students, a common practice is to compare teachers' grades with students' results in standardized tests administered at the national level – such as INVALSI (Dardanoni et al. 2009; Triventi 2020). This approach relies on estimating a grade equation model in which grades are compared with more "objective" assessment of students' competences such as standardized test scores (Dardanoni et al. 2009). In this way, it is possible to determine to what extent teachers consider students' competences when assessing them, and whether they consider other aspects such as ascriptive characteristics or socioemotional skills. To understand whether belonging to a specific students' profile is correlated to a higher or lower teacher assessment, controlling for students' characteristics, I rely on OLS regression models on students assigned grades in two subjects – language and mathematics. The basic linear grade equation for the two subjects (s) is:

$$Grade_{i_s} = \alpha + \delta(profile_{i_s}) + \beta_1(C_{i_s}) + \beta_2(tscore_{i_s}) + \beta_3(Z_{i_s}) + \varepsilon_{i_s} \quad (1)$$

Where $grade$ measures teacher assessment of student i in subject s, δ represents the total effect of students' profiles (socioemotional skills) on teacher assessment, C represents a vector of students' ascriptive characteristics (gender, migratory background, socioeconomic background), $tscore$ represents subject-specific competences (INVALSI test score), and Z represents a vector of other controls variables. Step-wise models control the bivariate association between students' socioemotional skills and teacher grades (model 1); model 2 includes ascriptive characteristics; model 3 includes INVALSI test scores in order to capture students' competences; model 4 includes other control variables in order to clean the estimated coefficients from other possible confounders. Finally, in order to investigates whether the relationship between students' ascriptive characteristics and teacher grades varies according to students' profiles, interaction terms are included between respectively, gender, ESCS and migratory background, and students' profiles.

Variables

To generate students' profiles, latent profile analysis is performed selecting, among the richness of PISA items and indices, those socioemotional and non-cognitive dimensions that are theoretically described as affecting teacher assessment. PISA provides a number of tested and validated indices[31] ("derived variables"), therefore in order to maximize the number of items and dimensions for the profiles' identification, 8 indices were selected among those present in the PISA dataset (OECD, forthcoming). The indexes measure such dimensions: competitiveness, fear of failure, eudaemonia (meaning in life), work mastery, learning goals, self-efficacy (resilience), cognitive flexibility/adaptability, and attitude toward school[32]. Table 4.1 provides information about the single items used in the creation of the indices.

The main independent variable is belonging to one of the identified students' profiles, as the result of the latent profile analysis. The main dependent variable is teacher assessment, or grades, in the two subjects of language and mathematics. Teacher grades range from 1 to 10, where 10 is the grade assigned to the highest academic performance, and 6 is the passing grade[33].

Students' gender, migratory background and socioeconomic background are the main control variables accounting for students' ascriptive characteristics. Gender is recoded as 0 if the student is male and as 1 if the student is female. Migratory background

[31] The PISA dataset provides indices derived from PISA 2018 student questionnaire. Indices are scaled using a two-parameter item-response model and values of the indices correspond to Warm likelihood estimates (WLE). For more information about the construction of indices, visit https://doi.org/10.1787/888934030838.

[32] Although the selected indexes may have different weights in explaining how teachers evaluate their students, (e.g., some characteristics may be more important than others), the idea behind this methodological choice was to analyze the relationship between the combination of different indexes among each other and teacher assigned grade, obtaining students' profiles as similar as possible to what teachers experience within the classroom context.

[33] Grades are computed as the average between teacher assessment in written exams and in oral exams, as reported in the school report at the end of the 1st semester of the relative academic year in which students performed the INVALSI test.

is recoded as 0 if the student is native, as 1 if the student is a 1st or 2nd generation immigrant. Finally, socioeconomic background (ESCS[34]) is recoded as a dummy variable in which 0 corresponds to the 1st and 2nd quartiles (lower ESCS) and 1 corresponds to the 3rd and 4th quartiles (higher ESCS).

Table 4.1: Selected indices for Latent Profile Analysis

Indices	*Items from student questionnaire (PISA 2018)*
1) Competitiveness	1) I enjoy working in situations involving competition with others; 2) It is important for me to perform better than other people on a task; 3) I try harder when I'm in competition with other people (response scale: a) strongly disagreed; b) disagreed; c) agreed; d) strongly agreed). Index positive values indicate higher competitiveness
2) Fear of failure	1) When I am failing, I worry about what others think of me; 2) When I am failing, I am afraid that I might not have enough talent; 3) When I am failing, this makes me doubt my plans for the future (response scale: a) strongly disagreed; b) disagreed; c) agreed; d) strongly agreed). Index positive values indicate higher fear of failure
3) Eudaemonia or meaning in life	1) My life has clear meaning or purpose; 2) I have discovered a satisfactory meaning in life; 3) I have a clear sense of what gives meaning to my life (response scale: a) strongly disagreed; b) disagreed; c) agreed; d) strongly agreed). Index positive values indicate higher eudaemonia
4) Work mastery	1) I find satisfaction in working as hard as I can; 2) Once I start a task, I persist until it is finished; 3) Part of the enjoyment I get from doing things is when I improve on my past performance (response scale: a) strongly disagreed; b) disagreed; c) agreed; d) strongly agreed). Index positive values indicate higher work mastery
5) Learning goals	1) My goal is to learn as much as possible; 2) My goal is to completely master the material presented in my classes; 3) My goal is to understand the content of my classes as thoroughly as possible (response scale: a) not at all true of me; b) slightly true of me; c) moderately true of me; d) very true of me; e) extremely true of me). Index positive values indicate higher learning goals

[34] ESCS is an index provided by INVALSI that measures students' economic, social and cultural status. It is a synthesis of three indicators: 1. Parental occupational status; 2. Parental level of education; 3. Possession of specific material assets.

6) Self-efficacy or resilience	1) I usually manage one way or another; 2) I feel proud that I have accomplished things; 3) I feel that I can handle many things at a time; 4) My belief in myself gets me through hard times; 5) When I'm in a difficult situation, I can usually find my way out of it (response scale: a) strongly disagreed; b) disagreed; c) agreed; d) strongly agreed). Index positive values indicate higher self-efficacy
7) Cognitive flexibility or adaptability	1) I can deal with unusual situations; 2) I can change my behaviour to meet the needs of new situations; 3) I can adapt to different situations even when under stress or pressure; 4) I can adapt easily to a new culture; 5) When encountering difficult situations with other people, I can think of a way to resolve the situation; 6) I am capable of overcoming my difficulties in interacting with people from other cultures (response scale: a) Very much like me; b) Mostly like me; c) Somewhat like me; d) Not much like me; e) Not at all like me). Index positive values indicate higher cognitive flexibility
8) School attitude	1) Trying hard at school will help me get a good job; 2) Trying hard at school will help me get into a good college; 3) Trying hard at school is important (response scale: a) strongly disagreed; b) disagreed; c) agreed; d) strongly agreed). Index positive values indicate positive attitude toward school

As indicators of students' subject-specific competences, the INVALSI test score performed in 10th grade in language and mathematics is included in order to allow the comparison of teacher grades for students with similar academic competences. INVALSI test score is a continuous variables used in its original scale, with mean of 200 and standard deviation of 40.

Final models control also for other confounders that might affect teacher judgment. Grade retention accounts for whether students repeated at least a year of schooling, and it is usually negatively associated with educational outcomes (García-Pérez et al. 2014). Geographical area accounts for the huge variability in teacher assessment in the Italian education system (Argentin & Triventi 2015). School track accounts for the possible differences in assessment between academic track, or lyceums, and vocational/technical tracks. Finally, academic year accounts for a possible cohort effect. Table A4.1 in the appendix contains information about the variables.

Results

Latent Profile Analysis

The first step is the identification of students' profiles. To identify students' profiles, I applied latent profile analysis (LPA) to the larger sample of students in the PISA dataset (10,680 observations) using the 8 indices. Latent Profile Analysis (LPA hereafter), similarly to Latent Class Analysis (McMutcheon 1987) and factor analysis, allows researchers to identify subgroups of an underlying categorical discrete latent variables. While latent class analysis allows the "characterization of a multidimensional discrete latent variable from a cross-classification of two or more observed categorical variables" (McMutcheon 1987:8), latent profile analysis is undertaken on continuous indicator variables (Williams and Kibowski, 2016). In this framework, indices have been normalized as continuous variables in order to facilitate the interpretation of results.

To fit a latent class model, it is necessary to specify the number of classes of the latent variable. Several goodness-of-fit tests were performed with different numbers of latent classes K (from 2 to 6 classes). Whereas the latent profile model with 6 classes was the optimal one according to model fit statistics (such as Bayesian and Akaike Information Criteria, see Table A4.2 in the appendix section), in the quest to strike a balance between highest model fit and model parsimony, the chosen option is the 4-class model. This allows to obtain a modest decrease in model-fit indicators while gaining a significant interpretability of results. Across all the different LPA specifications, one or more residual classes are identified. In the latent profile model with four classes, one residual class includes only 138 students, which are excluded from the analysis. The dropped group represented 1,36% of the final analytical sample and it is unlikely to be interpretable in a substantive way.

The conditional probabilities for each item within each latent class are analysed in order to understand the students' characteristics linked to the probability of belonging to a specific class, and to assign labels to the latent classes. The cross-case matrix (Miles & Huberman 1994) allows to substantially interpret the profiles identified through LPA, and to understand the socioemotional skills of students belonging to each profile and the differences between profiles (Table 4.2). The cross-case matrix reports sample means and profile means for each index. Colours are included for visually representing distances from the sample mean for each profile.

Table 4.2: Cross-case matrix with gradients indicating the distance from the mean for each index for the three profiles. Cross-case matrix refers to the analytical sample means (N = 6,504)

	Competi-tiveness	Fear of failure	Meaning in life	Work mastery	Learning goals	Self-efficacy	Cognitive flexibility	School attitude	%
Profile 1	0.462	0.498	0.379	0.499	0.352	0.440	0.467	0.527	27.64
Profile 2	0.588	0.533	0.515	0.778	0.558	0.561	0.531	0.756	53.7
Profile 3	0.751	0.442	0.761	0.917	0.748	0.825	0.704	0.825	18.66
Mean	0.581	0.513	0.514	0.728	0.534	0.569	0.540	0.708	

Note: Colors represent gradients, or distances from the mean for each index: yellow indicates a distance from the mean up to -/+ 0.05 points; orange indicates a distance from the mean up to – 0.15 points; red indicates a distance from the mean up to – 0.25 points; light green indicates a distance from the mean up to +0.15 points; green indicates a distance from the mean up to +0.25 points; dark green indicates a distance from the mean up to +0.5 points.

Gradients of green indicate a higher value compared to the sample mean for the specific index (positive socioemotional skills and higher fear of failure), while gradients of orange indicate a lower value compared to the sample mean for the specific index (negative socioemotional skills and lower fear of failure).

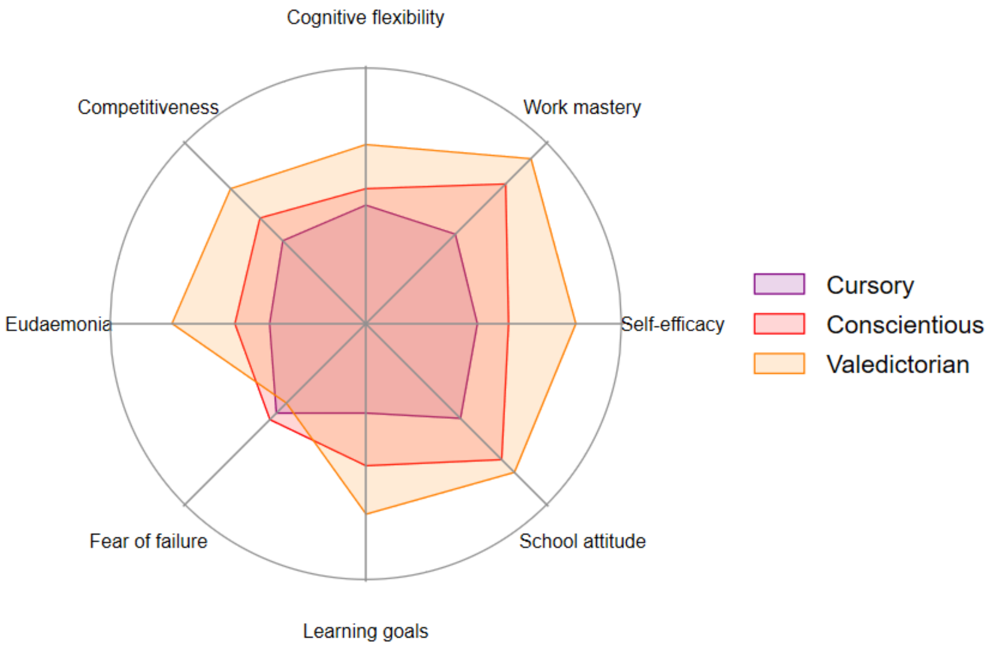
According to the distribution of the socioemotional skills selected, three profiles are identified; and to facilitate the interpretation of results, profiles are named as cursory, conscientious and valedictorian, respectively. "*Cursory students"* account for about 28% of the sample, and they generally show mean values below the average for each index. Work mastery, learning goals and attitude toward school are the lowest values, respectively 0.499, 0.352 and 0.527 in a scale that goes from 0 to 1. Interestingly, the lower values reflect items that specifically refer to the schooling/educational dimension (e.g., "my goal is to completely master the material presented in my classes", or "trying hard at school is important"). While only fear of failure aligns with the average, cursory students show overall poor socioemotional and noncognitive skills. This type of student seems to find no satisfaction in working hard or improving, and it has little interest in learning. 27,64% of the sample is represented in this profile.

The second student profile is represented by the "*conscientious students",* accounting for 53.7% of the sample. Conscientious students show average levels of each identified socioemotional skills. However, they show high levels of work mastery (0.778) and attitude toward schools (0.756).

The third student profile shows exceptionally high values for each socioemotional skills: this profile includes the *"valedictorian students"*. Valedictorian students, which represent 18,66% of the sample, show extremely high level of self-efficacy, or resilience, compared to the other profiles. In addition, they score above the average also with respect to all the other dimensions. Figure 4.1 illustrates the averages for the 8 selected socioemotional skills conditional on belonging to each of the three student profiles, where mean = 0 is the centre of the circle.

As highlighted in the cross-case matrix, radarplots clearly show the gradient in increasing average socioemotional skills from cursory, to conscientious, to valedictorian students. In summary, LPA allowed to explore the heterogeneity of students' socioemotional skills and school attitude as relational systems, and to highlight three different groups showing similar combination of such skills. Whereas previous research focused on central tendencies for single items, this approach allows to consider a number of different students' non-cognitive characteristics and apprise a operate a partitioning that, to the best of the author knowledge did not emerge in previous contributions.

Figure 4.1: Radarplots showing the profiles of the three identified classes: averages on the eight indices used in the Latent Profile Analysis (N = 6,504).



The Social Stratification of Student Profiles

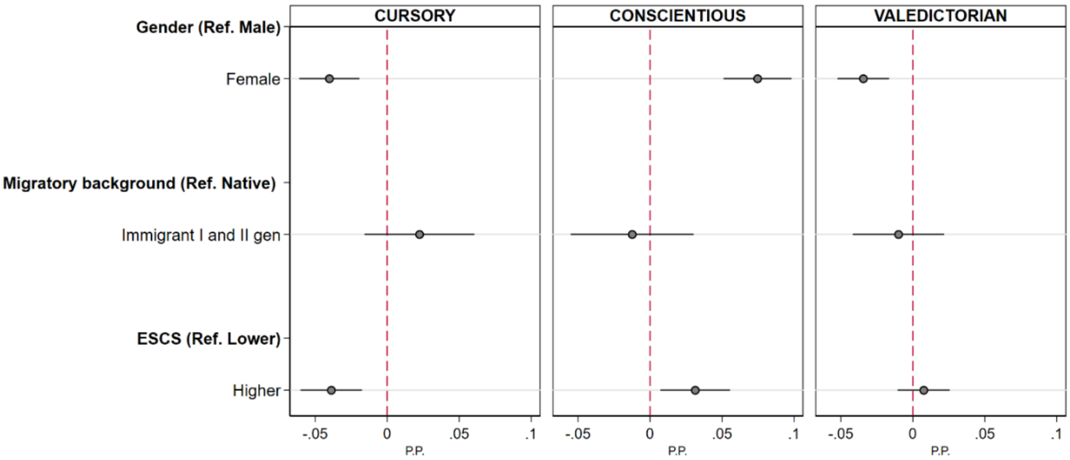
The second step of the analysis is to understand whether the probability of belonging in each of the identified student profile – cursory, conscientious or valedictorian – is different according to students' gender, socioeconomic background and migratory background. Figure 4.2 illustrates the result of a multinomial logistic regression model predicting the probability of belonging to each of the three classes (predicted probabilities) as a function of students' ascriptive characteristics.

Results show that students' gender has a statistically significant association with the profile membership. Female students have a higher probability than male students of belonging to the *conscientious* student type (+ 7.5 p.p.). In contrast, female students are less likely to belong to the *cursory* student type (- 4 p.p.) and to the *valedictorian* student type (-3.4 p.p.).

Migratory background does not have a significant statistical association with the membership to student profiles, even if it is possible to derive some patterns[35] indicating that immigrant students may be more likely to belong to the cursory type; while for the conscientious and the valedictorian profiles, coefficients for having a migratory background are very close to zero.

[35] Probably the small proportion of immigrant students in the sample (around 8 % counting together 1st and 2nd generation immigrants) does not allow to evidence significant relationship between migratory background and the probability of belonging to the cursory type of students.

Figure 4.2: Multinomial logistic regression predicting the probability of belonging to each student profile by students' gender, migratory background and socioeconomic background (ESCS); predicted probabilities and 95% confidence intervals (N = 6,504).



Note: coefficients for each model are controlled for all ascriptive characteristics. Robust standard errors.

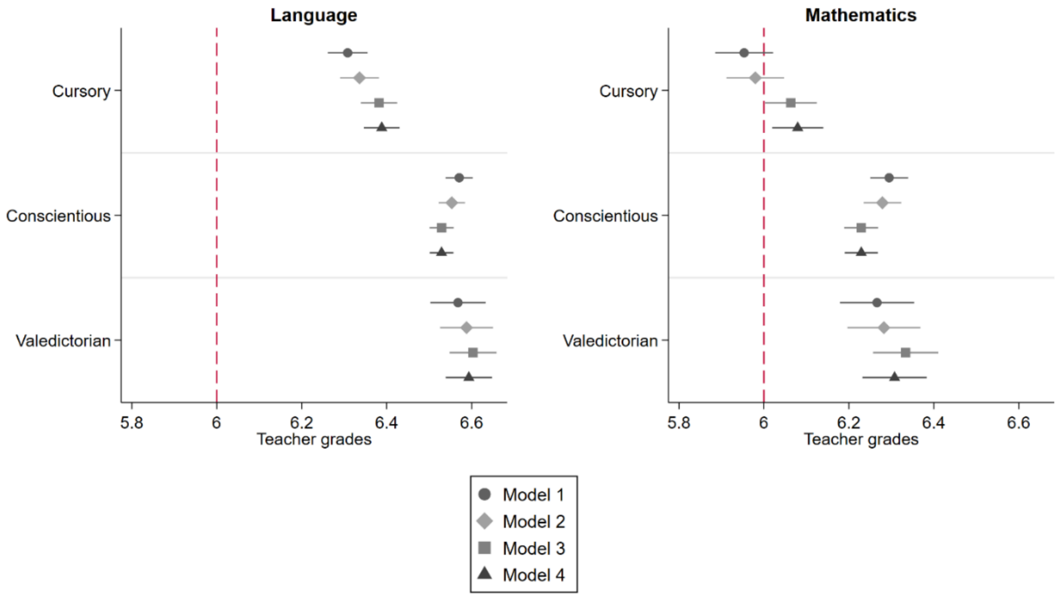
Looking at socioeconomic background, the results indicate that students with higher ESCS compared to student with lower ESCS have a higher probability to belong to the *conscientious* type of student (+ 3.1 p.p.) and a lower probability of belonging to the *cursory* type (- 3.8 p.p.), whereas with regard to the *valedictorian* type of students, socioeconomic background does not correlate with the probability of belonging to the group. Overall, the valedictorian type of student seems to be the most homogeneous with regard to migratory and socioeconomic background: only gender predicts the probability of belonging to this group. Boys are more likely to show exceptionally high levels in the selected socioemotional skills and non-cognitive traits.

Student Profiles and Teacher Assessment

The third step of the analysis is assessing whether belonging to a specific student profile is associated to an increase or a decrease in teacher assessed grades, over and above students' subject-specific competences and students' ascriptive characteristics. Figure 4.3 shows the predictive margins derived from different linear regression models for the two subjects – language and mathematics.

Looking at the linear predicted probabilities, the first noticeable result is that cursory students are assessed with significantly lower grades by their teachers compared to the conscientious and the valedictorian students. This holds in both subjects, even if in mathematics the gap between cursory students and other students is bigger. In the bivariate model (model 1), cursory students have an average grade of 6.3 in language and of 5.9 in mathematics, compared to the average grade of conscientious students (respectively 6.6 and 6.3) and of valedictorian students (respectively 6.6 and 6.3). Interestingly, conscientious students and valedictorian students show no differences in their average grades in the two subjects, even if students belonging to the valedictorian profile show more positive values in a number of socioemotional skills. The results hold also when controlling for students' sociodemographic characteristics such as gender, migratory background and socioeconomic background (model 2). More importantly, the difference in the average grade between the profiles is stable also when controlling for students' subject-specific competences (model 3) and for additional controls (model 4). This suggest that students' differences in socioemotional skills and soft skills determine differences in grades also when students have similar sociodemographic characteristics as well as similar academic competences.

Figure 4.3: Linear predicted probabilities derived from OLS models predicting teacher grade in Language and Mathematics in 10th grade; 95 % confidence intervals of students' profiles (N = 6,504).

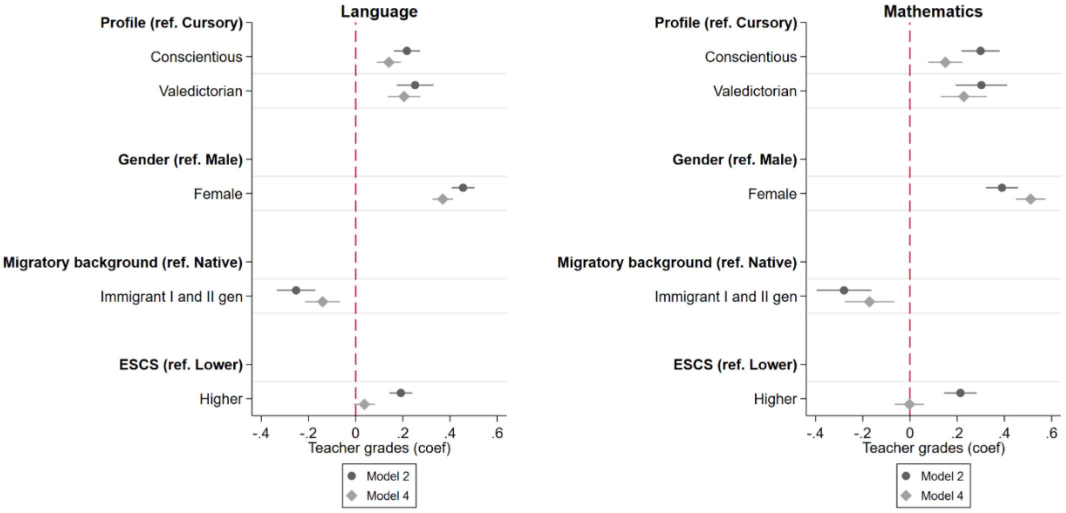


Note: Model 1 is the bivariate regression model predicting teacher grade by students' profiles; model 2 controls for students' ascriptive characteristics (gender, migratory background, socioeconomic background); model 3 controls additionally for students' subject-specific competences (INVALSI test score in Language and Mathematics); model 4 controls additionally for geographical area, school track (vocational, technical, lyceum), grade retention and academic year. Robust standard errors.

These results suggest that the negative impact of students' profiles on grades is noticeable only when students show levels of the selected socioemotional skills that fall below the sample average – meaning when students show particularly poor non-cognitive and socioemotional skills compared with their peers and classmates, such as students belonging to the cursory profile.

Focusing on students' ascriptive characteristics, Figure 4.4 represents coefficients for the grade predictors derived from OLS model 2 and model 4. Compared to model 2 – that includes controls only for students' profiles and sociodemographic characteristics – model 4, that includes all controls, shows that the differences in grade in both language and mathematics between cursory students and conscientious and valedictorian students diminishes. Figure 4.4 also shows that gender has a significant impact on grades, even when controlling for the rich set of socioemotional and non-cognitive skills included. This holds also including subject-specific competences. Indeed, looking at model 4, being female is associated with a higher grade both in language (+ 0.4) and in mathematics (+ 0.5). For mathematics, the differences between male and female grades increases once accounting for students' competences and other controls. Immigrant students seem penalized in term of grades respect to their native counterparts with similar characteristics, even if once academic competences are included, the gap become smaller (- 0.14 in language and -0.17 in mathematics). Finally, students with higher ESCS do not show significantly higher grades compared to students with lower ESCS when adding as controls to socioemotional skills and sociodemographic characteristics, also academic competences.

Figure 4.4: OLS models predicting teacher grade in Language and Mathematics in 10th grade; 95% confidence intervals of students' profiles, gender, migratory background and ESCS (N = 6,504).



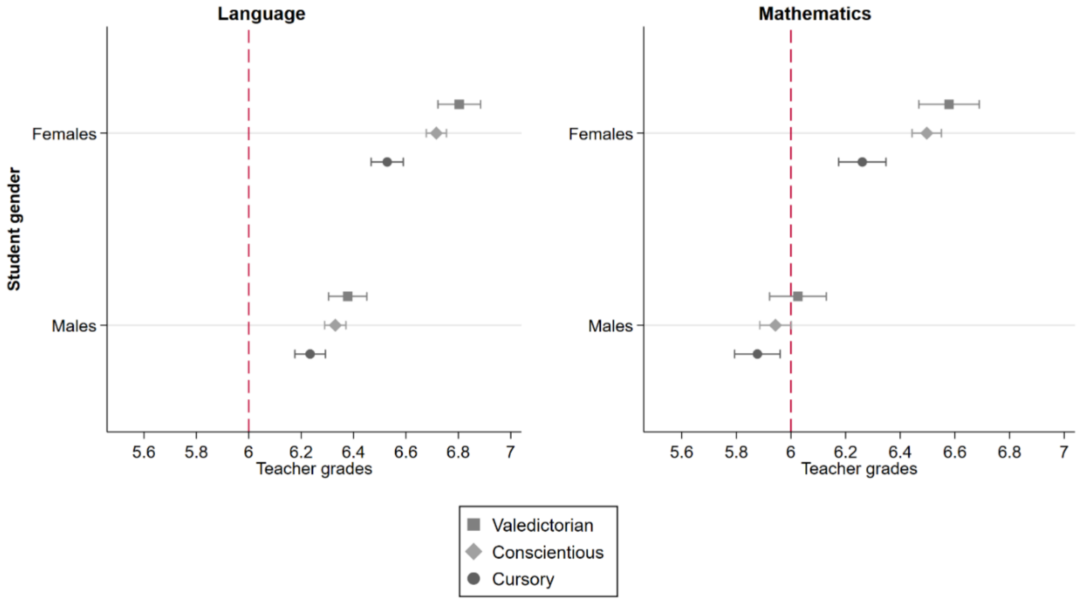
Note: Model 2 controls for students' profiles and students' ascriptive characteristics (gender, migratory background, socioeconomic background); model 4 controls additionally for students' subject-specific competences (INVALSI test score in Language and Mathematics), geographical area, school track (vocational, technical, lyceum), grade retention and academic year. Robust standard errors.

To understand whether the gap in teacher assessment between students' sociodemographic characteristics is similar across students belonging to different profiles, several OLS models including interaction terms are proposed. Figure 4.5, 4.6 and 4.7 show the interaction terms between students' ascriptive characteristics (respectively gender, migratory background, and ESCS) and students' profiles on teacher grades in the two subjects, controlling for students' subject-specific competences and all controls (model 4).

In Figure 4.5, is analysed the relationship between students' gender and teacher assessment according to students' profiles, compared to previous models that highlighted a higher average grade for female students controlling for students' socioemotional skills. Concerning teacher grade in Language, the advantage of being female rather than male in terms of grades is similar across students' profiles. In other words, belonging to a profile instead of another does not have any influence in increasing or decreasing the gender gap in grading for that profile[36]. Analysing teacher assessment in Mathematics, instead, the advantage of being female is slightly higher if students belong to the conscientious type of students ($p < 0.05$), therefore comparing conscientious girls and conscientious boys. Interestingly, being a male student belonging to the cursory and to the conscientious profiles makes a significative difference in terms of grade, since the predicted grade in mathematics fall below 6 – which is considered the passing mark in the Italian educational system.

[36] This is true considering 95% confidence intervals ($p < 0.05$).

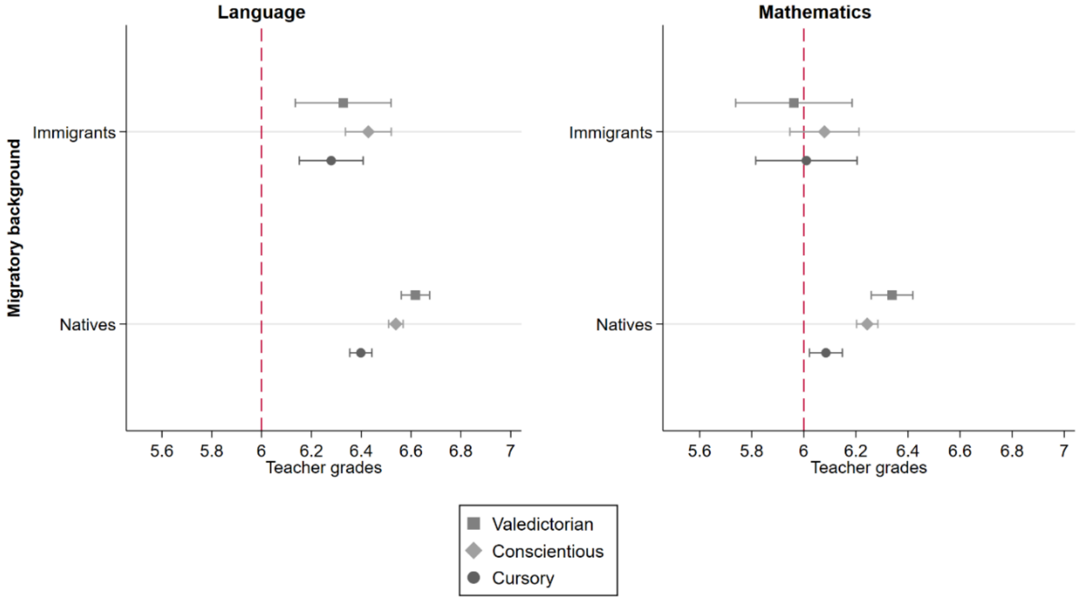
Figure 4.5: Linear predicted probabilities derived from OLS models predicting teacher grade in Language and Mathematics in 10th grade. Predictive margins for interaction terms between students' profiles and gender; 95% confidence intervals of students' profiles (N = 6,504).



Note: Coefficients are derived from model 4, that controls for subject-specific competences (INVALSI test score in Language and Mathematics), geographical area, school track (vocational, technical, lyceum), grade retention and academic year. Robust standard errors.

Figure 4.6 shows the relationship between students' migratory background and teacher assessment according to the three students' profiles. Results underline that the negative association between having an immigrant background and teacher grade in Language is similar across students with different profiles, suggesting that the relationship between migratory background and teacher assessment does not change according to students' socioemotional skills. Concerning teacher assessment in mathematics, instead, results suggest that immigrant students belonging to the valedictorian type are even more penalized in term of grades in Mathematics, since the gap in grading between native students and immigrant students becomes bigger when looking at valedictorian students ($p<0.05$). The predicted grade in mathematic for valedictorian immigrant students fall below the passing grade 6.

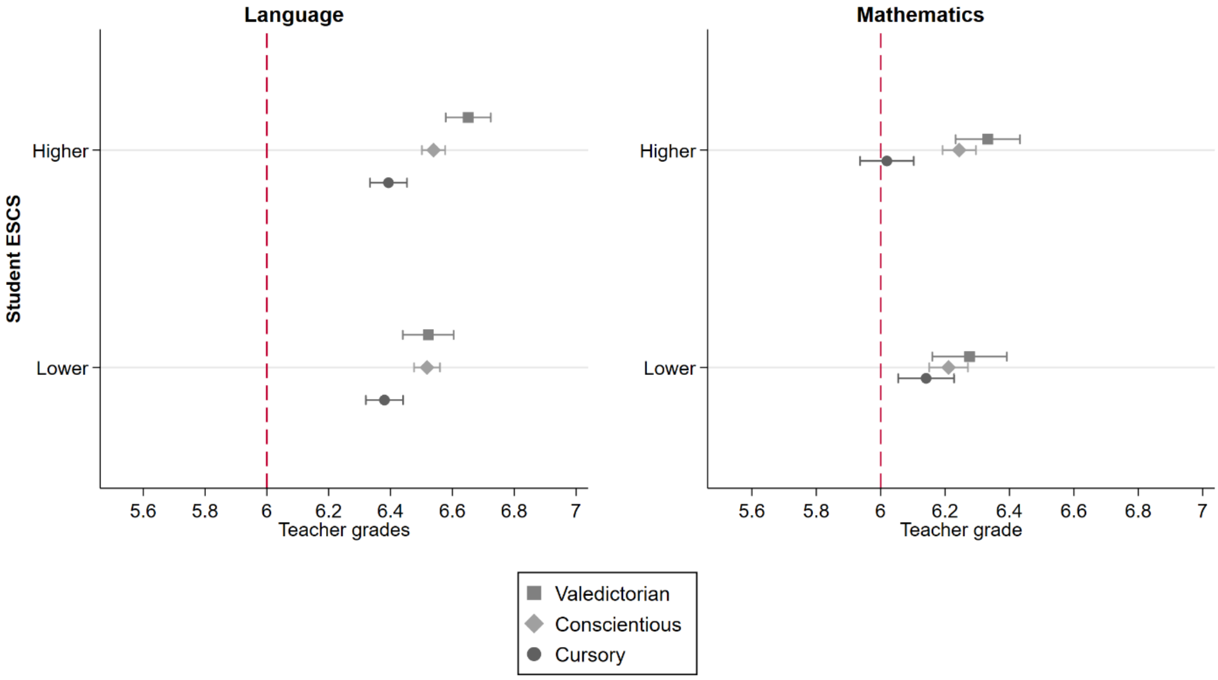
Figure 4.6: Linear predicted probabilities derived from OLS models predicting teacher grade in Language and Mathematics in 10th grade. Predictive margins for interaction terms between students' profiles and migratory background; 95% confidence intervals of students' profiles (N = 6,504).



Note: Coefficients are derived from model 4, that controls for subject-specific competences (INVALSI test score in Language and Mathematics), geographical area, school track (vocational, technical, lyceum), grade retention and academic year. Robust standard errors.

Concerning the interaction between students' profiles and students' ESCS (Figure 4.7), no significant patterns emerge if we consider as the outcome teacher grade in Language, meaning that there is no statistically significant association between students' socioeconomic background and teacher assessment, and this is similar across students' profile. Concerning teacher grade in Mathematics, however, results show that students with higher socioeconomic background are more advantaged when they belong to the conscientious type of student ($p<0.05$) compared to students with lower socioeconomic background that show conscientious socioemotional traits.

Figure 4.7: Linear predicted probabilities derived from OLS models predicting teacher grade in Language and Mathematics in 10th grade. Predictive margins for interaction terms between students' profiles and socioeconomic background (ESCS); 95% confidence intervals of students' profiles (N = 6,504).



Note: Coefficients are derived from model 4, that controls for subject-specific competences (INVALSI test score in Language and Mathematics), geographical area, school track (vocational, technical, lyceum), grade retention and academic year. Robust standard errors.

Robustness Checks

In this section, possible biases in the estimation are addressed. First, the huge loss of cases may arise when merging the two datasets because of the non-random selection of students according to observable characteristics. In order to correct the estimates, I create IPW (inverse probability weighting) performing binomial logistic regression models on the probability of being in the analytical sample, therefore of being matched, compared to the PISA sample (N = 10,174). The probability of being matched is performed on a number of students' observable covariates such as: migratory background, gender, region, school track, grade retention, cultural possession at home, highest parental education status, highest parental occupation status. Then I computed predicted probabilities from this logistic regression model, and weights are created as the inverse of the predicted probability.

Another possible bias in the estimation could stem from the selection of students into profiles according to some unobservable variables. In order to control for the selection into treatment, another set of models is proposed, including weight generated with the MMWS (marginal mean weighting through stratification) method. MMWS is a data pre-processing procedure that reweights a dataset to balance the observed pre-treatment characteristics across all treatment groups. The MMWS method removes the selection bias associated with the membership to a student profile, by equating the composition between groups. Unlike propensity score matching, the MMWS method is flexible for evaluating binary and multivalued treatments by approximating a completely randomized experiment (Linden, 2014), which is ideal in this setting in which the treatment variable is nominal. In this context, the observed pre-treatment characteristics selected are: gender, migratory background, socioeconomic background, school track, geographical area and INVALSI test score in 8th grade[37] in the two subjects – measured two years before the INVALSI test in 10th grade.

Another possible bias could arise considering the fact that some students performed the INVALSI test score one year before or one year after the administration of

[37] Another advantage of this dataset is that it is possible to match previous INVALSI test scores thank to the SIDI code, the student identifier. This allows to follow students throughout their school career. Unfortunately, this implies a further loss of cases due to attrition from 8th grade to 10th grade. The sample, including information about INVALSI test scores in the two subjects in 8th grade, is composed by 6,150 students (354 students are missing). The mmws weighting leads to another 48 observations dropped because of lack of common support. The sample is composed by 6,102 students.

the PISA students' questionnaire measuring socioemotional traits and school attitude. Indeed, only students matched with the INVALSI 2017-2018 perform the test the same period. The assumption that socioemotional skills and personality traits measured by the selected indices are quite stable over time (Briley & Tucker-Drob 2014) may not hold considering the adolescents sample (Morris et al. 2021).

In the appendix (Figure A4.1) I present a comparison between five different estimated models: the first model controls for the selected variables for the final model (model 4, N = 6,504); the second model includes inverse probability weighting (N = 6,504); the third model includes weights generated with mmws method (N = 6,102); the fourth model shows the analysis performed only on the subsample of matched students in 10th grade in the academic year 2017-2018 (N = 6,144); the fifth model shows results including mmws on the subsample of matched students in 10th grade in the academic year 2017-2018 (N = 5,854).

Results show no considerable differences between the five models. Moreover, since the inclusion of subject-specific competences in 8th grade and consequently of mmws leads to a significant loss of cases, as selecting only students matched from academic year 2017-2018, the final model does not include these specifications in order to maximize the sample numerosity.

Conclusion and Discussion

This chapter aims at bridging a gap in the literature of teacher grade determinants, analysing teacher grades as a function of students' academic competences – measured via INVALSI standardized test scores – together with students' ascriptive characteristics –

gender, migratory background and socioeconomic background – and students' socioemotional skills and non-cognitive factors. This has been accomplished by assessing the non-random distribution of non-cognitive factors among the student sample, and considering the interdependence of socioemotional skills among each other.

Results indicate that Italian 15-years-old students can be partitioned in three profiles according to the within-group similarities in the distribution of non-cognitive and socioemotional skills, together with schooling attitudes. The three student profiles have been labelled, according to the manifested non-cognitive dimensions, as: cursory students, conscientious students, valedictorian students.

Student gender significantly predicts the membership to each of the three profiles: while female students are more likely to belong to the conscientious profiles, male students are more likely to belong to the cursory and valedictorian profiles. Concerning migration background, results show no significant correlations between being a 1st or 2nd generation immigrant students and the belonging to different profiles. However, considering the small proportion of immigrant students in the sample, is it possible to hypothesize that the latter may be more likely to belong to the cursory type of students and less likely to belong to the conscientious type. Finally, students with a higher socioeconomic status are less likely to belong to the cursory type of students, and more likely to belong to the conscientious type of students. Concerning the valedictorian profile, neither migratory background nor socioeconomic background significantly predict the belonging to this profile.

Advancing in the analysis, results show also that student profiles significantly predict teacher assigned grades in both Mathematics and Language. This result holds when students' subject specific competences, as well as students' ascriptive characteristics, are included in the model. Conscientious students and valedictorian types of students are assessed by their teachers between 0.15 and 0.22 points higher in Language and Mathematics compared to the cursory type of students, even when including all control variables. In addition, results show that female students and native students are assessed with significantly higher grade by their teachers, even when controlling for students' profiles – as a rich set of socioemotional and non-cognitive skills, as well as for students' subject-specific competences (see model 4 in Table A4.3 and A4.4 in appendix).

When looking at the interaction terms, results are different concerning the two subjects. For grades in Language, the association between students' characteristics such migratory background or socioeconomic background and teacher grades does not change according to students' profiles. Instead, looking at grades in Mathematics, when accounting for students' gender, in addition to the gender grading gap favouring girls, conscientious female students exhibit an additional advantage in teacher assessment (compared to conscientious boys). Similarly, conscientious high ESCS students show the same additional advantage in teacher assessment in Mathematics, compared to conscientious low ESCS students.

These results lead to several conclusions. First of all, non-cognitive factors are not randomly distributed across the student sample. Indeed, some groups are more likely to display specific features and socioemotional skills traits that may be highly rewarded in the educational and schooling context by their teachers and educators. This may lead to a systematic group advantage that adds up to, and partially confirm, teacher expectation bias.

Interestingly, the disadvantage in terms of grades according to students' socioemotional skills is noticeable only when students show below average non-cognitive skills. Indeed, there are no significant differences in grades between the conscientious type of students and the valedictorian type of students, even if the valedictorian students show incredibly high levels of non-cognitive skills and socioemotional traits. This is true for both Language and Mathematics. This may suggest that, within the classroom context, teachers may pick useful information for their assessment only when students display particularly bad schooling attitudes and socioemotional skills. In other words, what may make a real difference for 10th grade students, or for adolescents more generally, is showing or having poor attitudes rather than brilliant socioemotional skills. Unfortunately, the dataset contains no information concerning the classroom composition. Indeed, this result may be due to the impossibility to control for the profiles' composition of the schools/classrooms: valedictorian students may stand in a class with average low socioemotional skills, and vice-versa. Moreover, also looking at the same association in different grades may lead to different conclusions.

Focusing on the difference between Language and Mathematics, a relevant result is that the relationship between socioemotional skills and teacher assessment differs across the two subjects: this contradicts the common belief that teachers' evaluations in mathematics' assignments may be less prone to teacher's subjectivity. The difference across the two subjects may be also explained by the teachers' autonomy in the Italian grading system. Indeed, at all levels of the educational systems, teachers have a great deal of freedom in deciding the type of exams students should take (e.g., oral test or written test, multiple-choice questions or open-ended questions etc.), the test frequency and, above all, the evaluation criteria.

The fact that the relationship between student profiles and teacher grades in Language is pretty similar independently of students' characteristics suggests a role played by socioemotional skills and schooling attitudes that goes beyond student ascriptive characteristics. However, looking at grades in Mathematics, results from the interaction terms indicate also that the relationship between students' ascriptive characteristics and teacher assessment may change according to students' profiles. Different socioemotional skills seem to be important when considering different characteristics. Concerning migratory background, only belonging to the valedictorian type seems to be less advantageous for immigrants. Valedictorian immigrant students are indeed penalized by their teachers compared to valedictorian native students in Mathematics. A possible interpretation for this result may derive from the fact that immigrant students displaying particularly positive schooling attitudes and socioemotional skills may collide with teachers' negative expectations about immigrant students (Alesina et al. 2018), which may in turn result as an over-penalization. This may not be the case for Language, since teachers tend to have a positive grading bias that may help disadvantaged students to cope with difficulties linked to their non-native status in Language (Alesina et al. 2018). However, misleading interpretations may arise since 1st and 2nd generation immigrants are considered together due to low numerosity of immigrant students.

Looking at gender, belonging to the conscientious type seems to be more advantageous for girls rather than for boys regarding grades in Mathematics. A possible explanation could arise from the consideration that teachers may unconsciously push girls with "average" socioemotional skills, trying to balance with a good grade the negative expectations related to girls as a stigmatized group concerning math ability (Lievore &

Triventi 2022). Finally, students with higher socioeconomic background belonging to the conscientious type are graded more generously than conscientious students with lower socioeconomic background. This may be explained considering the cultural reproduction framework (Cole & Mendick 2006): students with higher ESCS may be more capable to display in the classroom context positive socioemotional skills that are subsequently rewarded in terms of grades by their teachers. Contrarily to valedictorian students, who probably manage to demonstrate their high levels of socioemotional skills no matter the social status, low ESCS conscientious students need to work more in order to actually show positive skills such as engagement, participation or self-control.

This study presents some limitations. First, since the PISA sample includes randomly selected students across Italian schools, it is not possible to account for robust effects clustered at the classroom or at the school level. Second, the analysis is based on students' self-reported items, therefore the results may be biased by differential item functioning. That is, if students have different proclivities in answering specific items according to their belonging to specific sociodemographic groups (e.g., by gender), it may affect both the results concerning the stratification of students' profiles and the final estimates. Even if previous studies found no strong differential item functioning effects for gender in PISA test results (see Khorramdel et al. 2020), it is still not clear whether this may affect group averages responses on socioemotional skills and attitude toward school items.

In conclusion, the strong association between the identified students' profiles and teacher assessment in each subject confirm the idea according to which students' profiles may capture the interplay between socioemotional skills and schooling attitude, and may come closer to what teachers experience within the classroom context when evaluating students. It is important to underline that the scope of this chapter was assessing a total effect of students' socioemotional skills on teacher grades, without considering the possible mechanisms explaining such total effect. Future studies may consider exploring such explanatory mechanisms, for example considering actual behaviours in the classroom linked to students' non-cognitive skills (e.g., homework behaviour, truancy, compliance with deadlines).

This chapter, overcoming the study of single correlations between students' indicators and teacher assessment, suggests a possible new approach for understanding the determinants of teacher grades. This has important policy implications if we consider the number of important life-outcomes linked to teacher grades, such as college admission, school drop-out, earning and so on. This approach advocates that in order to gain a more comprehensive understanding of how teacher assign grades, the teacher grade equation should always include, together with student academic ability, also socioemotional skills, ascriptive characteristics and possibly their simultaneous interplay.

Appendix Chapter 4

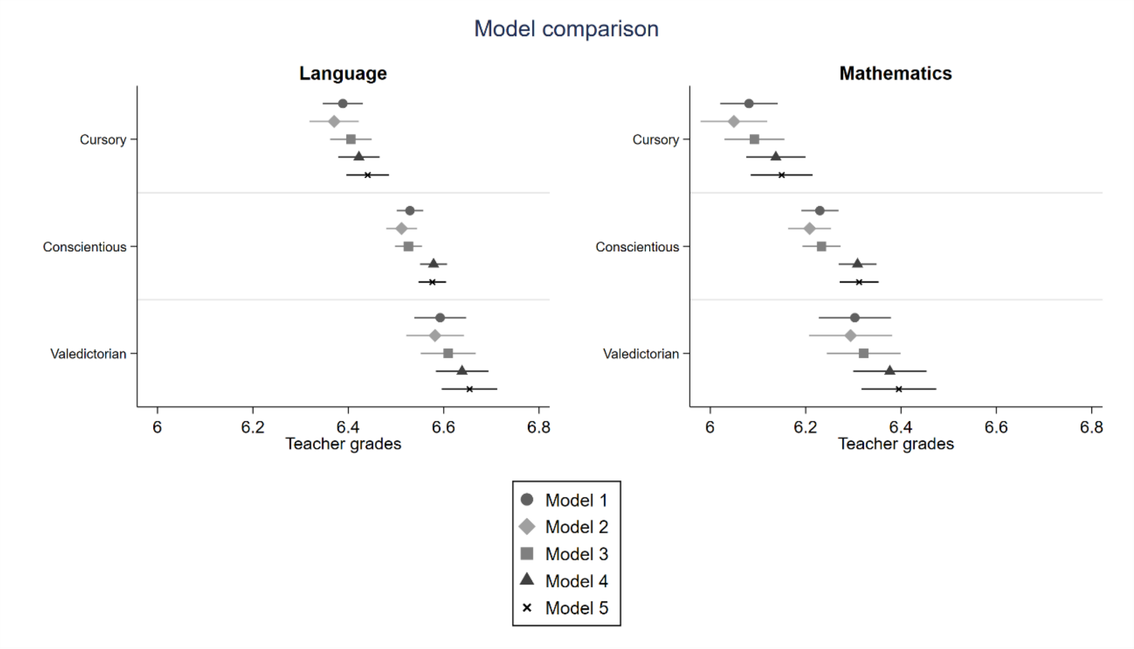
Table A4.1: Description of variables (N = 6,504)

Variables	Description
Dependent variable	
Teacher grade in Language Teacher grade in Mathematics	Teacher grade in mid-school report (*pagella primo quadrimestre).* Average grade between oral examinations and written examinations in both subjects. Scale 1-10 (where 6 is the passing mark)
Independent variable	
Students' profiles	Students' profiles derived from Latent Profile Analysis recoded in 0 = Cursory students; 1 = Conscientious students; 2 = Valedictorian students
Main control variables	
INVALSI test score in Language INVALSI test score in Mathematics	Continuous variable measuring subject-specific competences in both subjects in 10th grade. Scores obtained by students in the INVALSI standardized tests. The scores have mean 200 and standard deviation 40
Gender	Recoded as 0 = male and 1 = female
Socioeconomic background	Index provided from INVALSI that measures students' economic, social and cultural status. It is a synthesis of three indicators: 1. Parental occupational status (HISEI); 2. Parental level of education (PARED); 3. Possession of specific material assets (HOMEPOS). Recoded as 0 = Lower ESCS if students are in the 1st and 2nd quartile and 1 = Higher ESCS if students are in the 3rd and 4th quartile.
Migratory background	Recoded as 0 = Natives and 1 = 1st and 2nd generation immigrants
Other control variables	
Geographical area	Recoded as 0 = North-West; 1 = North-East; 2 = Centre; 3 = South; 4 = Isles (as defined by ISTAT, national istitute of statistics)
School track	Recoded as 0 = Vocational schools; 1 = Techical schools; 2 = Lyceums
Grade retention	Recoded as 0 = Never; 1 = At least once (including also primary school)
Academic year	Recoded as 0 = A.Y. 2017-2018; A.Y. 2018-2019

Table A4.2: Latent Profile Analysis and model fit statistics (AIC and BIC) fitting from 2 to 6 classes.

Model	N	Ll (model)	df	AIC	BIC
2 profiles	10,680	11231.53	25	-22413.06	-22231.16
3 profiles	10,680	12228.34	34	-24388.68	-24141.29
4 profiles	10,680	12797.92	43	-25509.85	-25196.98
5 profiles	10,680	13288.88	52	-26473.75	-26095.4
6 profiles	10,680	13422	61	-26722.01	-26278.16

Figure A4.1: Model comparison predicting grade in Language and in Mathematics in grade 10. Linear predicted probabilities from OLS models predicting teacher grades by students' profiles; 95% confidence intervals. (Model 1: N = 6,504; Model 2: N = 6,504. Model 3: N = 6,102. Model 4: N = 6,144. Model 5: N = 5,854).



Note: Model 1 controls for: students' ascriptive characteristics (gender, migratory background, socioeconomic background), subject-specific competences (INVALSI test score in grade 10 Language and Mathematics, respectively), geographical area, school track (vocational, technical, lyceum), grade retention and academic years. Model 2 includes inverse probability weighting. Model 3 includes marginal mean weighting through stratification. Model 4 includes all controls and is performed on the subsample of students matched in the academic year 2017-2018. Model 5 includes all controls, marginal mean weighting trough stratification and is performed on the subsample of students matched in the academic year 2017-2018. Weights are generated through propensity scores for the treatments, that are estimated with a multinomial logistic regression model that includes: students' gender, socioeconomic background, ethnic background, school track, geographical area, INVALSI test score in language and mathematics in 8th grade.

Table A4.3: OLS models predicting teacher grade in Language. Robust standard errors in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

Teacher grade in Language	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Student profile (Ref. Cursory)							
Conscientious	0.263***	0.217***	0.148***	0.141***	0.096***	0.140***	0.138***
	(0.029)	(0.028)	(0.026)	(0.026)	(0.036)	(0.027)	(0.037)
Valedictorian	0.260***	0.252***	0.222***	0.204***	0.142***	0.218***	0.140***
	(0.041)	(0.039)	(0.036)	(0.035)	(0.048)	(0.037)	(0.052)
Female (Ref. Male)		0.455***	0.377***	0.365***	0.290***	0.366***	0.365***
		(0.024)	(0.022)	(0.023)	(0.043)	(0.023)	(0.023)
Immigrant (Ref. Native)		-0.252***	-0.114***	-0.141***	-0.143***	-0.121*	-0.141***
		(0.042)	(0.039)	(0.038)	(0.038)	(0.069)	(0.038)
Higher ESCS (Ref. Lower)		0.192***	0.042*	0.036	0.036	0.036	0.013
		(0.025)	(0.023)	(0.023)	(0.023)	(0.023)	(0.043)
INVALSI score Language grade 10			0.425***	0.411***	0.410***	0.410***	0.410***
			(0.012)	(0.013)	(0.013)	(0.013)	(0.013)
Geographical area (Ref. North-West)							
North-East				0.099***	0.097***	0.099***	0.099***
				(0.036)	(0.035)	(0.036)	(0.036)
Center				0.012	0.011	0.012	0.012
				(0.035)	(0.035)	(0.035)	(0.035)
South				-0.106***	-0.109***	-0.105**	-0.106***
				(0.041)	(0.041)	(0.041)	(0.041)
Isles				0.049	0.047	0.048	0.047
				(0.039)	(0.039)	(0.039)	(0.039)
School track (Ref. Vocational)							
Technical				-0.091**	-0.092**	-0.093**	-0.092**
				(0.040)	(0.039)	(0.040)	(0.040)
Lyceums				-0.140***	-0.140***	-0.142***	-0.139***
				(0.040)	(0.040)	(0.040)	(0.040)
Grade retention at least once (Ref. Never)				-0.257***	-0.257***	-0.259***	-0.257***
				(0.096)	(0.096)	(0.096)	(0.096)
Academic year 2018-19 (Ref. 2017-18)				-0.775***	-0.779***	-0.773***	-0.776***
				(0.057)	(0.057)	(0.057)	(0.057)
Female # Conscientious					0.091*		
					(0.051)		
Female # Valedictorian					0.132*		
					(0.070)		
Immigrant # Conscientious						0.011	
						(0.084)	
Immigrant # Valedictorian						-0.170	
						(0.123)	
Higher ESCS # Conscientious							0.007
							(0.051)
Higher ESCS # Valedictorian							0.117*
							(0.070)
Constant	6.308***	6.019***	6.143***	6.287***	6.324***	6.285***	6.298***
	(0.024)	(0.029)	(0.027)	(0.049)	(0.051)	(0.049)	(0.052)
Observations	6,504	6,504	6,504	6,504	6,504	6,504	6,504
R-squared	0.013	0.077	0.230	0.258	0.259	0.258	0.259

Table A4.4: OLS models predicting teacher grade in Mathematics. Robust standard errors in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

Teacher grade in Mathematics	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Student profile (Ref. Cursory)							
Conscientious	0.341***	0.299***	0.166***	0.149***	0.065	0.156***	0.071
	(0.042)	(0.041)	(0.037)	(0.037)	(0.051)	(0.038)	(0.053)
Valedictorian	0.313***	0.303***	0.270***	0.222***	0.141**	0.248***	0.126*
	(0.057)	(0.056)	(0.050)	(0.049)	(0.068)	(0.052)	(0.073)
Female (Ref. Male)		0.390***	0.525***	0.489***	0.362***	0.490***	0.489***
		(0.035)	(0.031)	(0.032)	(0.062)	(0.032)	(0.032)
Immigrant (Ref. Native)		-0.279***	-0.156***	-0.177***	-0.181***	-0.090	-0.179***
		(0.059)	(0.055)	(0.053)	(0.053)	(0.103)	(0.053)
Higher ESCS (Ref. Lower)		0.214***	-0.005	-0.006	-0.006	-0.006	-0.124**
		(0.035)	(0.032)	(0.032)	(0.032)	(0.032)	(0.061)
INVALSI score Mathematics grade 10			0.628***	0.620***	0.619***	0.619***	0.619***
			(0.016)	(0.018)	(0.018)	(0.018)	(0.018)
Geographical area (Ref. North-West)							
North-East				0.065	0.062	0.066	0.067
				(0.049)	(0.049)	(0.049)	(0.049)
Center				0.050	0.049	0.051	0.051
				(0.049)	(0.049)	(0.049)	(0.049)
South				-0.082	-0.087	-0.081	-0.080
				(0.056)	(0.056)	(0.056)	(0.056)
Isles				0.114**	0.113**	0.113**	0.114**
				(0.056)	(0.056)	(0.056)	(0.056)
School track (Ref. Vocational)							
Technical				-0.291***	-0.290***	-0.291***	-0.289***
				(0.058)	(0.057)	(0.058)	(0.058)
Lyceums				-0.315***	-0.315***	-0.316***	-0.314***
				(0.057)	(0.057)	(0.057)	(0.057)
Grade retention at least once (Ref. Never)				-0.413***	-0.412***	-0.416***	-0.413***
				(0.131)	(0.131)	(0.130)	(0.130)
Academic year 2018-19 (Ref. 2017-18)				-1.318***	-1.323***	-1.314***	-1.319***
				(0.067)	(0.067)	(0.067)	(0.067)
Female # Conscientious					**0.169****		
					(0.073)		
Female # Valedictorian					0.172*		
					(0.098)		
Immigrant # Conscientious						-0.073	
						(0.125)	
Immigrant # Valedictorian						**-0.298****	
						(0.159)	
Higher ESCS # Conscientious							**0.151****
							(0.073)
Higher ESCS # Valedictorian							0.183*
							(0.099)
Constant	5.953***	5.687***	5.731***	6.064***	6.127***	6.055***	6.122***
	(0.035)	(0.043)	(0.039)	(0.067)	(0.071)	(0.068)	(0.072)
Observations	6,504	6,504	6,504	6,504	6,504	6,504	6,504
R-squared	0.010	0.038	0.218	0.258	0.259	0.258	0.259

Conclusions

This dissertation aims at deepening the understanding of a central topic in sociology, the reproduction of educational inequalities, focusing on the role that teachers may play within the classroom context in shaping inequalities according to students' ascriptive characteristics, such as gender, ethnic origin and socioeconomic background.

The starting point is the development of a comprehensive theoretical framework that accounts for teacher effects in the broader framework attributed to Boudon (1974) aiming to explain the variety of mechanisms underlying educational inequalities. In relation to primary and secondary effects – respectively, social-background differences in academic performance, and social-background differences in educational choices, the role of teachers as external influences is unclear, even if they are known to be very influential concerning students' development of their competences and students' allocation in different tracks. They also play an important role in the reproduction of educational inequalities, since they may be biased according to students' ascriptive characteristics, and this can have fundamental consequences for students' educational outcomes and trajectories. Through their grading practices and their recommendations, they have a strong, and mostly unconscious, role in determining which students have academic potential and which do not.

In this dissertation, I propose a theorization of the role of teachers taking advantage of the definition of tertiary effects (Esser 2016), that add to the primary-and-secondary effects model and account for how teachers may be biased according to students' ascriptive characteristics and how this may have long-term consequences for students' educational career. Tertiary effects are meant to capture inequalities in educational attainment due to the active role of teachers when showing different expectations according to students' different (social) backgrounds (Thys 2018).

The analysed context is the Italian educational system at different grades of compulsory schooling. Italy is an interesting country for studying teacher grading practices and its consequences, because the high level of educational inequalities in compulsory education and the heterogeneous territorial divides (also in terms of school resources) is accompanied by no formal restrictions linked to teacher grades or recommendations and a great deal of autonomy in teachers' duties. The analyses rely on the INVALSI dataset, which is a rich population sample of students that aims at assessing students' competences through a standardized tests along compulsory schooling, gathering information on their grades, on their socioeconomic characteristics, as well as on their teachers. This important feature makes it well-suited for investigating the role of teachers in the reproduction of educational inequalities in the Italian system, allowing the use of Italian student-teacher matched data.

This dissertation proposed three different empirical chapter making use of the INVALSI-SNV data, with the aim of providing new evidence about teacher effects, the role of teachers in the reproduction of educational inequalities, and the consequences of teacher grading practices. Although not all the empirical chapters refer directly to the tertiary effects definition provided in Chapter 1, they all aim at shading light on the micro-mechanisms underlying the complex effects that teachers and grading practices may have on students' educational outcomes. The goal of Chapter 1 is therefore providing a comprehensive theoretical understanding about the role of teachers, that may be thought as a broader framework in which the single empirical chapters are embedded, with their more specific research questions. The methodological connection between the empirical chapters is the use of the grade equation model, in which teacher grades are expressed as a function as a variable identifying the group of interest – such as gender, ethnicity or social status – plus an “objective” measure of student academic ability. The comparison between the two measures, teacher grade from one side, and student score in a standardized test from the other side, may provide the extent to which teachers are likely to reward or penalize students from different social groups, or may serve as a way of measuring specific grading practices.

An example of the latter is provided in chapter 2, which does not refer directly to tertiary effects, but it is embedded in the potential effect that teachers may have in determining students’ educational career. Indeed, is an example of how the grade equation model – that is, the comparison between teacher grades and a more “objective” measure of students’ ability – may be used in order to study teacher effects from a different perspective. The attention is shifted from teachers’ expectation bias to the impact that different grading practices may have on students’ belonging to different social groups. The goal of the second chapter is to analyse the impact of having a strict rather than a generous teacher, with regards to later students’ competences in language and mathematics as well as to their probability to choose an academic track. After creating a measure of teacher grading standards, I rely on an instrumental variable approach in order to determine whether higher grading standards measured at the end of primary schools have an impact on students’ educational outcomes measured later in time. The results of Chapter 2 demonstrate that students with a stricter teacher in primary schools have higher performance in both Language and Mathematics in 8th and 10th grade and are more likely to be enrolled in the academic track in 10th grade, and this effect is stable among students belonging to different social groups.

After having focused on the teacher effect with an educational trajectory perspective, the third and the fourth chapters aim at investigating deeper some mechanisms concerning tertiary effects and teacher grading bias. The first focus in on the role of contextual factors, while the second focus is on students' socioemotional skills.

More specifically, the third chapter provides empirical evidence of the gender grading gap in Italian upper secondary schools, while analysing the role of teachers' characteristics, classroom composition and school type in shaping gender grading mismatch. Results show that, while teachers are more likely to grade female students with higher grades in two subjects, this premium in grade is stable even when accounting for a number of contextual factors regarding teachers' characteristics, classroom composition and type of upper secondary school. Unfortunately, this gender grading gap is not accountable for students' attitudes and behaviours, since there is no information on the INVALSI-SNV dataset in this regard. However, I wanted to account for other educational signals that may determine teacher expectations besides students' ascriptive characteristics, such as students' behaviour in classroom, attitude towards school, effort, participation and socioemotional skills.

Consequently, for Chapter 4 I construct a novel dataset merging the INVALSI data and the PISA data, with the aim of analysing teacher judgments accounting for students' socioemotional skills and attitude toward school, net of students' sociodemographic characteristics and students' academic performance measured through standardized test scores. I created students' profiles according to their similarities in the distribution of their socioemotional and non-cognitive skills in order to understand if they are related to teacher grades. The student profile with lower socioemotional skills is associated with lower grades, over and above gender, socioeconomic background and migratory background. While being a female student and a native student is always associated with higher grades, also when controlling for socioemotional skills. However, while for Language the relationship between ascriptive characteristics and grade is not moderated by student's socioemotional skills, in Mathematics specific social groups belonging to specific student profiles might receive boosted grades – adding an additional advantage.

Some general conclusions can be derived from the results of the empirical analysis about tertiary effects, and teacher effects more generally, and the role of teachers in the Italian educational system in shaping educational inequalities in compulsory schooling. Some micro-mechanisms involving teachers may contribute to the reproduction of social inequalities in the education systems related to gender, ethnic background and social class. However, these seem related more to teachers' expectation related to students' characteristics rather than to the classroom or the school context. This suggests that i) referring to tertiary effects as school effects may be misleading and it may not capture the entire set of mechanisms underlying what are commonly called "school effects"; ii) focusing only on primary and secondary effects, therefore on parents' and students' intentions and choices, may be a limitation for quantitative educational studies in the understanding of the complexity of inequalities reproduction in education. Moreover, introducing a new perspective on teacher effects, focusing not only on class-based inequalities but also on differentiations according to students' gender and ethnicity may help broadening our knowledge on the topic. This, in accordance with Argentin and Pavolini (2020), must be addressed also considering the macro-institutional settings and characteristics of the educational system under analysis, such as levels, tracking, or teacher allocation, in order to fully grasp the action played by schools, principals and teachers.

The studies presented in this dissertation show that the role of teachers in affecting both students' academic performance and educational choices is strong, even in an educational system such as the Italian one where there is formally a high level of centralization and low level of school autonomy, and in which teachers' grade and recommendations are not binding in accessing specific school tracks. Students' ascriptive characteristics, and more specifically students' gender, are still great determinants of students' evaluations, over and above their actual academic competences. This is true even when considering in the equation several determinant socio-emotional skills and non-cognitive traits, that, even if stratified according to students' characteristics, are not capturing the entire variation between academic competences and teacher grades. The present studies suggest also that, even if most of the literature and previous studies focus on the role of social class, it may be less relevant in the study of teacher effects compared to other students' characteristics.

From a social intervention perspective, some final thoughts may also be derived from these empirical studies. First, the availability of large dataset linking information of students with information of teachers and schools permits to investigate more deeply how educational inequalities are reproduced within schools (Argentin & Pavolini 2020). It suggests following this direction in terms of data accessibility, in order to allow researchers, policy evaluators and educational experts to design appropriate targeted interventions. Secondly, these studies suggest that intervening in order to level possible mechanisms in educational inequalities through teacher practices may be relatively cost effective, for example manipulating teacher grading standards which are influential in determining both academic competences and school track choices. Finally, part of the solution for fostering equal opportunities may lie at the schools and teachers' level: teachers should be aware of their (mostly) unconscious role in the reproduction of social, ethnic and gender inequalities in the Italian education system, and of their potential role in reducing inequalities (Geven et al. 2018).

References

Abbiati, G., Argentin, G., Gerosa, T. (2017): Different Teachers for Different Students? Evidence on Teacher-Student Matching and its Consequences in the Italian Case, in *Journal of Economic Policy*, 33(1):13-58.

Alesina, A., Carlana, M. La Ferrara, E., Pinotti, P. (2018): Revealing Stereotypes: Evidence from Immigrants in Schools, in *NBER,* Working Paper: 25333.

Alexander, K.L., Entwisle, D.R., Olson, L.S. (2001): Schools, Achievement, and Inequality: A Seasonal Perspective, in *Educational Evaluation and Policy Analysis*, 23(2):171-191.

Altonji, J. G., Elder E. T., Taber R. C. (2005): Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools, in *Journal of Political Economy*, 113(1):151-184.

Ammermueller, A., Dolton, P. (2006): Pupil-Teacher Interaction Effects on Scholastic Outcomes in England and the USA, in *ZEW* Discussion Papers 06–60.

Anders, Y., McElvany, N., Baumert, J. (2010): Perception of Learning Relevant Student Characteristics at School Transition. How Differentiated are Teacher Judgements? in *Der Übergang von der Grundschule in die weiterführende Schule: Leistungsgerechtigkeit und Regionale, Soziale und Ethnisch-kulturelle Disparitäten*, in Maaz, K., Baumert, J., Gresch, C., McElvany, N. (Eds): 313–330, Bonn: Bundesministerium für Bildung und Forschung, Referat Bildungsforschung.

Andrei, F., Mancini, G., Mazzoni, E., Russo, P.M., Baldaro, B. (2015): Social Status and its Link with Personality Dimensions, Trait Emotional Intelligence, and Scholastic Achievement in Children and Early Adolescents, in *Learning and Individual Differences*, 42:97–105.

Angelo, C.S. (2014): Is there a Bias Towards Girls in Non Anonymous Evaluation?, Working Paper.

Argentin, G., Barbieri, G., Barone, C. (2017): Origini Sociali, Consiglio Orientativo e Iscrizione al Liceo: Un'Analisi Basata sui Dati dell'Anagrafe Studenti, in *Social Policies*, 1:53-73.

Argentin, G., Pavolini, E. (2020): How Schools Directly Contribute to the Reproduction of Social Inequalities. Evidence of Tertiary Effects, Taken from Italian Research, in *Politiche Sociali*, 1:149-176.

Argentin, G., Triventi, M. (2015): The North-South Divide in School Grading Standards: New Evidence from National Assessments of the Italian Student Population, in *Italian Journal of Sociology of Education*, 7(2):157-185.

Aucejo, E.M., Coatez, P., Fruehwirthx, C.J., Kelly, S., Mozenterk, Z. (2021): Teacher Effectiveness and Classroom Composition: Understanding Match Effects in the Classroom, in *The Economic Journal*, 132(648):3047-3064.

Azzolini, D., Barone, C. (2012): Between Old and New Forms of Inequality: The Educational Attainment of Immigrant Children in Italian Upper Secondary Education, in *Italian Review of Sociology*.

Azzolini, D., Barone, C. (2012): Between Old and New Forms of Inequality: The Educational Attainment of Immigrant Children in Italian Upper Secondary Education, in *Italian Review of Sociology*.

Azzolini, D., Mantovani, D., Santagati, M. (2018): Four Emerging Traditions in Immigrant Education Studies, in Stevens, P.A., Dworkin G.A. (Eds): *The Palgrave Handbook of Race and Ethnic Inequalities in Education,* UK, Palgrave.

Azzolini, D., Mantovani, D., Santagati, M. (2018): Four Emerging Traditions in Immigrant Education Studies, in Stevens, P.A., Dworkin G.A. (Eds): The Palgrave Handbook of Race and Ethnic Inequalities in Education, UK, Palgrave.

Babad, E. (2009): Teaching and Nonverbal Behavior in the Classroom, in Saha, L.J., Dworkin, A.G. (Eds.): *International Handbook of Research on Teachers and Teaching*, Boston, MA: Springer, 817–827.

Babad, E.Y., Inbar, J., Rosenthal, R. (1982): Pygmalion, Galatea, and the Golem: Investigations of Biased and Unbiased Teachers, in *Journal of Educational Psychology*, 74:459–474.

Ballarino, G., Panichella, N., Triventi, M. (2014): School Expansion and Uneven Modernization. Comparing Educational Inequality in Northern and Southern Italy, in *Mobility*, 36:69-86.

Barg, K. (2012): The Influence of Students' Social Background and Parental Involvement on Teachers' School Track Choices: Reasons and Consequences, in *European Sociological Review*, 29(3):565-579.

Barone, C. (2011): Some Things Never Change: Gender Segregation in Higher Education across Eight Nations and Three Decades, in *Sociology of Education*, 84:157-176.

Barone, C., Ruggera, L. (2018): Educational Equalization Stalled? Trends in Inequality of Educational Opportunity Between 1930 and 1980 Across 26 European Nations, in *European Societies*, 20(1):1-25.

Basilio, J.R., Almeida, A.M.F. (2018): Teacher Employment Contracts and Student Performance, in *Revista Brasileira de Educação,* 23(0):1–23.

Baudson, T.G., Fischbach, A., Preckel, F. (2016): Teacher Judgments as Measures of Children's Cognitive Ability: A Multilevel Analysis, in *Learning and Individual Differences*, 52:148-156.

Becker, D. (2013): The Impact of Teachers' Expectations on Students' Educational Opportunities in the Life Course: An Empirical Test of a Subjective Expected Utility Explanation, in *Rationality and Society*, 25(4):422-469.

Becker, G.S., Hubbard, W.H., Murphy, K.M. (2010): Explaining the Worldwide Boom in Higher Education of Women, in *Journal of Human Capital,* 4(3):203–41.

Becker, W., Rosen, S. (1992): The Learning Effect of Assessment and Evaluation in High School, in *Economics of Education Review*, 11(2):107–118.

Behaghel, L., Crepon, B., Le Barbanchon., T. (2015): Unintended Effects of Anonymous Resumes, in *American Economic Journal: Applied Economics*, 7(3):1–27.

Ben-Peretz, M. (2001): The Impossible Role of Teacher Educators in a Changing World, in *Journal of Teacher Education*, 52(1):48-56.

Bernstein, B. (1975): *Class, Codes and Control, Volume 3: Towards a Theory of Educational Transmissions*. London: Routledge and Kegan Paul.

Bernstein, B. (1996): *Pedagogy, Symbolic Control and Identity*. London: Taylor and Francis.

Bernstein, B. (2000): *Pedagogy, Symbolic Control and Identity: Theory, Research Critique*. Lanham, MD: Rowman & Littlefield

Betts, J.R. (1997): Do Grading Standards Affect the Incentive to Learn?, University of California at San Diego, Department of Economics, Discussion Paper 97-22.

Betts, J.R. (1998): The impact of educational standards on the level and distribution of earnings, in *American Economic Review*, 88:266-275.

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

Birkelund, J.F., van de Werfhorst, H.G. (2022): Long-term Labor Market Returns to Upper Secondary School Track Choice: Leveraging Idiosyncratic Variation in Peers' Choices, in *Social Science Research*, 102:102629.

Blatchford, P., Bassett, P., Brown, P. (2011): Examining the Effect of Class Size on Classroom Engagement and Teacher-pupil Interaction: Differences in Relation to Pupil Prior Attainment and Primary vs. Secondary Schools, in *Learning and Instruction*, 21:715-730.

Blatchford, P., Bassett, P., Goldstein, H., Martin, C. (2003): Are Class Size Differences Related to Pupils' Educational Progress and Classroom Processes? Finding from the Institute of Education Class Size Study of children aged 5-7 Years, in *British Educational Research Journal*, 29(5):709-730.

Blatchford, P., Moriarty, V., Edmonds, S., Martin, C. (2002): Relationships Between Class Size and Teaching: A Multi-Method Analysis of English Infant Schools, in *American Educational Research Journal*, 39(1):101-132.

Boerma, I.E., Mol, S.E., Jolles, J. (2016): Teacher Perceptions Affect Boys' and Girls' Reading Motivation Differently, in *Reading Psychology*, 37:547–569

Bonesrønning, H. (1999): The Variation in Teachers' Grading Practices: Causes and Consequences, in *Economics of Education Review*, 18:89–105.

Bonesrønning, H. (2004): Do the Teachers' Grading Practices Affect Student Achievement?, in *Education Economics*, 12(2):151-167.

Bonizzoni, P., Romito, M., Cavallo, C. (2016): Teachers' Guidance, Family Participation and Track Choice: The Educational Disadvantage of Immigrant Students in Italy, in *British Journal of Sociology of Education*, 37(5):702-720.

Bonner, S.M., Chen P.P. (2019): Chapter 3: The Composition of Grades: Cognitive and Noncognitive Factors, in Guskey T.R., Brookhart S.M. (Eds): *What We Know about Grading: What Works, What Doesn't, and What's Next*, ASCD, Alexandia (VA).

Boone, S., Van Houtte, M. (2012): Social Inequalities in Educational Choice at the Transition From Primary to Secondary Education: A Matter of Rational Calcuation?, in *Culture and Education*, 91(5).

Boone, S., Van Houtte, M. (2013): In Search of the Mechanisms Conducive to Class Differentials in Educational Choice: A Mixed Method Research, in *The Sociological Review*, 61:549-572.

Boone, S., Van Houtte, M. (2013): Why are Teacher Recommendations at the Transition from Primary to Secondary Education Socially Biased? A Mixed-methods Research, in *British Journal of Sociology of Education,* 34(1):20–38

Bordalo, P., Coffman, K., Gennaioli, N., Shleifer, A. (2016): Stereotypes, in *The Quarterly Journal of Economics,* 131(4):1753–1794.

Borghans L., Golsteyn B.H.H., Heckman J.J., Humphries J.E. (2016): What Grades and Achievement Tests Measure, in *PNAS*, 113(47):13354–13359.

Borghans, L., Golsteyn, B. H., Heckman, J., and Humphries, J. E. (2011): Identification Problems in Personality Psychology, in *Personality and Individual Differences*, 51:315–320.

Borgna, C., Struffolino, E. (2017): Pushed or Pulled? Girls and Boys Facing Early School Leaving Risk in Italy, in *Social Science Research*, 61:298-313.

Boudon, R. (1974): Basic Mechanisms Generating Inequality of Educational Opportunity, in Terry N.C. (eds): Education, Opportunity, and Social Inequality. Changing Prospects in Western Society, in Wiley Series in Urban Research

Bound, J., Brown, C., Mathiowetz, N. (2001): Measurement Error in Survey Data, in Heckman, J., Leamer, E. (Eds): *Handbook of Econometrics,* Chapter 59.

Bourdieu P., Passeron C. (1990): Reproduction in Education, Society, Culture, London: Sage.

Bourdieu, P. (1974): Cultural Reproduction and Social Reproduction, in Brown, R. (Eds): *Knowledge, Education and Cultural Change: Papers in the Sociology of Education*, London, Tavistock.

Bowers, A.J. (2011): What's in a Grade? The Multidimensional Nature of What Teacher-assigned Grades Assess in High School, in *Educational Research and Evaluation*, 17:141–159.

Bowers, A.J., Sprott, R., Taff, S. (2013): Do We Know Who Will Drop Out? A Review of the Predictors of Dropping Out of High School: Precision, Sensitivity and Specificity, in *The High School Journal*, 96:77–100.

Bozzano, M. (2012): Assessing Gender Inequality among Italian Regions: The Italian Gender Gap Index, in *Quaderni di Dipartimento*, 174, Università degli Studi di Pavia, Dipartimento di Economia Politica e Metodi Quantitativi (EPMQ).

Bracci, E. (2009): Autonomy, Responsibility and Accountability in the Italian School System, in *Critical Perspective on Accounting,* 20(3):293–312.

Breen, R., Goldthorpe, J.H. (1997): Explaining Educational Differentials: Towards a Formal Rational Action Theory, in *Rationality and Society*, 9:275.

Breen, R., Luijkx, R., Müller, W., Pollak, R. (2009): Nonpersistent Inequality in Educational Attainment: Evidence from Eight European Countries, in *American Journal of Sociology*, 114(5):1475-1521.

Brennan, D.J. (2008): University Student Anonymity in the Summative Assessment of Written Work, in *Higher Education Research & Development*, 27(1):43–54.

Briley, D.A., Tucker-Drob, E.M. (2014): Genetic and Environmental Continuity in Personality Development: a Meta-Analysis, in *Psychol. Bull.*, 140(5):1303–31.

Brookhart, S.M. (1991): Grading Practices and Validity, in *Educational Measurement: Issues and Practice*, 10(1):35–36.

Brookhart, S.M. (1993): Teachers' Grading Practices: Meaning and Values, in *Journal of Educational Measurement*, 30:123–142.

Brookhart, S.M., Guskey, T.R., Bowers, A.J., McMillan, J.H., Smith, J.K., Smith, L.F., Stevens, M.T., Welsh, M.J. (2016): A Century of Grading Research: Meaning and Value in the Most Common Educational Measure, in *Review of Educational Research*, 86(4):803-848.

Brophy, J., Good, T. (1970): Teachers' Communication of Differential Expectations for Children's Classroom Performance: Some Behavioral Data, in *Journal of Educational Psychology*, 61(5):365−374.

Brophy, J.E. (1983): Research on the Self-fulfilling Prophecy and Teacher Expectations, in *Journal of Educational Psychology*, 75(5):631-661.

Brophy, J.E., Good, T. (1974): Teacher-student Relationships: Causes and Consequences, New York: Holt.

Bruhwiler, C., Blatchford, P. (2011): Effects of Class Size and Adaptive Teaching Competency on Classroom Processes and Academic Outcome, in *Learning and Instruction*, 21:95-108.

Buchmann, C., DiPrete, T.A. (2006): The Growing Female Advantage in College Completion: The Role of Family Background and Academic Achievement, in *American Sociological Review,* 71(4):515–541.

Burgess, S., Greaves, E. (2013): Test Scores, Subjective Assessment, and Stereotyping of Ethnic Minorities, in *Journal of Labor Economics*, 31(3):535–576.

Bygren, M. (2020): Biased grades? Changes in Grading after a Blinding of Examinations Reform, in *Assessment & Evaluation in Higher Education*, 45(2):292-303.

Card, D., Payne, A.A. Payne. (2017): *High School Choices and the Gender Gap in Stem*. Technical report, National Bureau of Economic Research.

Carlana, M. (2019): Implicit Stereotypes: Evidence from Teachers' Gender Bias, in *The Quarterly Journal of Economics,* 134(3):1163–1224.

Carroll, J.B. (1993): *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*, Cambridge University Press.

Ceci, S.J., Williams, W.M. (Eds) (2007): *Why Aren't More Women in Science?: Top Researchers Debate the Evidence*, American Psychological Association.

Checchi, D. (2004): Da Dove Vengono le Competenze Scolastiche?, in *Stato e Mercato,* 72:413–453.

Chen, P.P., Bonner, S.M. (2017): Teachers' Beliefs About Grading Practices and a Constructivist Approach to Teaching, in *Educational Assessment,* 22(1):18–34.

Cherry, T.L., Ellis, L.V. (2005): Does Rank-Order Grading Improve Student Performance? Evidence From a Classroom Experiment, in *International Review of Economic Education,* 4(1):9–19.

Cheryan, S., Ziegler, S.A., Plaut, V.C., Meltzoff, A.N. (2014): Designing Classrooms to Maximize Student Achievement, in *Policy Insights from the Behavioral and Brain Sciences*, 1(1):4–12.

Chetty R., Friedman J. N., Rockoff J. E. (2014): Measuring the Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood, in *American Economic Review*, 104(9):2633–79.

Chowdhury, F. (2018): Grade Inflation: Causes, Consequences and Cure, in *Journal of Education and Learning*, 7(6).

Chulkov, D.V. (2006): Student Response to Grading Incentives: Evidence from College Economics Courses, in *Journal of Instructional Psychology*, 33(3):206-211.

Clark, D.C. (1969): Competition for Grades and Graduate Student Performance, in *The Journal of Educational Research*, 62:351–354.

Clotfelter, C.T., Ladd, H.F., Vigdor, J.L. (2007): How and Why do Teacher Credentials Matter for Student Achievement?, *NBER* Working Paper No.12828, Retrieved from National Bureau of Economic Research website.

Cochran-Smith, M. (2000): Teacher Education at the Turn of the Century, in *Journal of Teacher Education*, 51(3):163-165.

Cole, M., Mendick, H. (2006): Education and Society: Issues and Explanations in the Sociology of Education, in *British Journal of Sociology of Education*, 27(1):111-123.

Conard, M. (2006): Aptitude is not Enough: How Personality and Behavior Predict Academic Performance, in *Journal of Research in Personality*, 40:339-346

Contini, D., Triventi, M. (2016): Chapter 18: Between Formal Openness and Stratification in Secondary Education: Implications for Social Inequalities in Italy, in Blossfeld, H-P., Buchholz, S., Skopek, J. and Triventi, M. (Eds): *Models of Secondary Education and Social Inequality*, Social and Political Science Collection, Cheltenham, UK; Edward Elgar 305-322.

Cooper, H.M., Good, T.L. (1983): *Pygmalion Grows Up: Studies in the Expectation Communication Process,* New York: Longman.

Cooper, H.M., Tom, D.Y.H. (1984): Teacher Expectation Research: A Review with Implications for Classroom Instruction, in *The Elementary School Journal,* 85(1):77–89.

Cornwell, C., Mustard, D.B., Van Parys, J. (2013): Noncognitive Skills and the Gender Disparities in Test Scores and Teacher Assessments: Evidence from Primary School, in *Journal of Human Resources*, 48(1):236-264.

Correa, H., Gruver, G.W. (1987): Teacher-Student Interaction: a Game Theoretic Extension of the Economic Theory of Education, in *Mathematical Social Science*, 13:19-47.

Correll, S.J, Benard, S. (2006): Biased Estimators? Comparing Status and Statistical Theories of Gender Discrimination. In: Thye S.R., Lawler E.J. (eds): Social Psychology of the Workplace, Vol 23, Advances in Group Processes. New York: Elsevier, pp. 89–116.

Costrell, R.M. (1994): A Simple Model of Educational Standards, in *The American Economic Review*, 84(4):956–971

Dardanoni, V., Modica, S., Pennisi, A. (2009): Grading Across Schools, in *The BE Journal of Economic Analysis and Policy*, 9(1):1-16.

Darley, J., Fazio, R. (1980): Expectancy Confirmation Processes Arising in the Social Interaction Sequence, in *American Psychologist*, 35:867−881.

de Zeeuw, E.L., van Beijsterveldt, C.E.M., Glasner, T.J., Bartels, M., de Geus, E.J.C., Boomsma, D.I. (2014): Do Children Perform and Behave Better at School When Taught by Same-Gender Teachers?, in *Learning and Individual Differences*, 36:152-156.

Dee, T.S. (2004): The Race Connection: Are Teachers More Effective with Students who Share Their Ethnicity?, in *Education Next*, 2:52–59.

Dee, T.S. (2005): A Teacher Like Me. Does Race, Ethnicity, or Gender Matter?, in *The American Economic Review,* 95:158–165.

Dee, T.S. (2007): Teachers and the Gender Gaps in Student Achievement, in *Journal of Human Resources*, 42:528–554.

DeVries, J.M., Rathmann, K., Gebhardt, M. (2018). How Does Social Behavior Relate to Both Grades and Achievement Scores?, in *Frontiers in Psychology*, 9:857

Di Liberto, A., Casula, L., Pau, S. (2021): Grading Practices, Gender Bias and Educational Outcomes: Evidence from Italy, in *Education Economics,* https://doi.org/10.1080/09645292.2021.2004999

Donnelly, M. (2018): Inequalities in Higher Education: Applying the Sociology of Basil Bernstein, in *Sociology*, 52(2):316–332.

Downey, D., Yuan, A.S.V. (2005): Sex Differences in School Performance During High School: Puzzling Patterns and Possible Explanations, in *Sociological Quarterly,* 46(2):299–321.

Dumais, S.A. (2002): Cultural Capital, Gender, and School Success: The Role of Habitus, in *Sociology of Education,* 75:44–68.

Eccles, J.S., Wong, C.A., Peck, S.C. (2006): Ethnicity as a Social Context for the Development of African-American Adolescents, in *Journal of School Psychology*, 44(5):407–426.

Egalite, A.J., Kisida, B., Winters, M.A. (2015): Representation in the Classroom: The Effect of Own-Race Teachers on Student Achievement, in *Economics of Education Review*, 45:44-52.

Egalite, A.J., Mills, J.N., Greene J.P. (2016): The Softer Side of Learning: Measuring Students' Non-Cognitive Skills, in *Improving Schools*, 19(1):27–40.

Ehrenberg, R.G., Brewer, D.J., Gamoran, A., Willms, J.D. (2001): Class Size and Student Achievement, in *Psychological Science in the Public Interest*, 2(1):1-30.

Ehrenberg, R.G., Goldhaber, D.D., Brewer, D.J. (1995): Do Teachers' Race, Gender and Ethnicity Matter? Evidence from the National Educational Longitudinal Study of 1988, in *Industrial and Labor Relations Review*, 48(3):54.

Elwert, F., Winship, C. (2014): Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable, in *Annual Review of Sociology*, 40(1):31-53.

Emanuelsson, I., Fischbein, S. (1986): Vive la Difference? A Study on Sex and Schooling, in *Scandinavian Journal of Educational Research,* 30:71–84.

Enzi, B. (2015): Gender Differentials in Test scores and Teacher Assessments: Evidence from Germany, Working Paper.

Esser, H. (2016): The Model of Ability Tracking. Theoretical Expectations and Empirical Findings on How Educational Systems Impact on Educational Success and Inequality, in Blossfeld, H.P., Buchholz, S., Skopek, J. Triventi, M. (Eds.): *Models of Secondary Education and Social Inequality: An International Comparison*, Cheltenham-Northampton: Edward Elgar Publishing.

Esser, H., Relikowski, I. (2015): Is Ability Tracking (Really) Responsible for Educational Inequalities in Achievement? A Comparison between the Country States Bavaria and Hesse in Germany, IZA DP n. 9082.

Facchinello, L. (2020): Short- and Long-run Effects of Early Grades, Available at SSRN: https://ssrn.com/abstract=2966571.

Falch, T., Naper, L.R. (2013): Educational Evaluation Schemes and Gender Gaps in Student Achievement, in *Economics of Educational Review,* 36:12–25.

Fallan, L., Opstad, L. (2012): Attitude towards Study Effort Responde to Higher Grading Standards: Do Gender and Personality Distinctions Matter?, in *Journal of Education and Learning*, 1(2):179-187.

Farkas, G., Grobe, R.P. Sheehan, D., Shuan, Y. (1990): Cultural Resources and School Success: Gender, Ethnicity, and Poverty Groups within an Urban School District, in *American Sociological Review,* 27(4):127–142.

Farrington, C.A., Roderick, M., Allensworth, E., Nagaoka, J., Keyes, T.S., Johnson, D. W. (2012): *Teaching Adolescents to Become Learners. The Role of Noncognitive Factors in Shaping School Performance: A Critical Literature Review.* Chicago: University of Chicago Consortium on Chicago School Research

Figlio, D.N., Lucas, M.E. (2004): Do High Grading Standards Affect Student Performance?, in *Journal of Public Economics*, 88:1815-1834.

Finefter-Rosenbluh, I., Levinson, M. (2015): What is Wrong with Grade Inflation (if Anything)?, in *Philosophical Inquiry in Education*, 23(1):3-21.

Finn, J.D., Achilles, C.M. (1999): Tennessee’s Class Size Study: Findings, Implications, Misconceptions, in *Educational Evaluation and Policy Analysis*, 21(2):97-109.

Fleming, N.D. (1999): Biases in Marking Students' Written Work: Quality?, in Brown, S., Glasner, A. (Eds): *Assessment Matters in Higher Education: Choosing and Using Diverse Approaches*, Buckingham, Open University Press.

Fletcher, J.M., Wolfe, B. (2016): The Importance of Family Income in the Formation and Evolution of Non-Cognitive Skills in Childhood, in *Economics of Education Review*, 54:143-154.

Frost, M.B. (2007): Texas Students' College Expectations: Does High School Racial Composition Matter?, in *Sociology of Education*, 80(1):43-65.

Fryer, R.G., Levitt, S. (2010): An Empirical Analysis of the Gender Gap in Mathematics, in *American Economic Journal: Applied Economics,* 2:210–40.

Furnham, A., Zhang, J., Chamoro, T. (2006): The Relationship Between Psychometric and Self-Estimated Intelligence, Creativity, Personality and Academic Achievement, in *Imagination, Cognition and Personality,* 25(2):119-145.

García-Pérez, J.I., Hidalgo-Hidalgo, M., Robles-Zurita, J.A. (2014): Does Grade Retention Affect Students’ Achievement? Some Evidence from Spain, in *Applied Economics*, 46(12):1373-1392.

Gentrup, S., Rjosk, C. (2018): Pygmalion and the Gender Gap: Do Teacher Expectations Contribute to Differences in Achievement between Boys and Girls at the Beginning of Schooling?, in *Educational Research and Evaluation*, 24(3-5): 295-323.

Gerbino, M., Zuffianò, A., Eisenberg, N., Castellani, V., Luengo Kanacri, B. P., Pastorelli, C. (2018): Adolescents’ Prosocial Behavior Predicts Good Grades Beyond Intelligence and Personality Traits, in *Journal of Personality*, 86:247–260.

Gershenson, S. (2016): Linking Teacher Quality, Student Attendance, and Student Achievement, in *Education Finance and Policy,* 11(2):125–49.

Geven, S., Batruch, A., van de Werfhorst, H. (2018): *Inequality in Teacher Judgements, Expectations and Track Recommendations: A Review Study*, report commissioned by the Dutch Ministry of Education, Culture and Sciences.

Geven, S., Batruch, A., van de Werfhorst, H. (2018): *Inequality in Teacher Judgements, Expectations and Track Recommendations: A Review Study*, report commissioned by the Dutch Ministry of Education, Culture and Sciences.

Gilbert, M.C., Musu-Gillette, L.E., Woolley, M.E., Karabenick, S.A., Strutchens, M.E., Martin, W.G. (2014): Student Perceptions of the Classroom Environment: Relations to Motivation and Achievement in Mathematics, in *Learning Environments Research*, 17:287–304.

Gillen-O' Neel, C., Ruble, D., Fuligni, A. (2011): Ethnic Stigma, Academic Anxiety, and Instrinsic Motivation in Middle Childhood, in *Child Development*, 82(5):1470–1485.

Glock, S., Kleen, H. (2017): Gender and Student Misbehavior: Evidence from Implicit and Explicit Measures, in *Teaching and Teacher Education,* 67:93–103.

Glock, S., Krolak-Schwerdt, S. (2014): Stereotype Activation versus Application: How Teachers Process and Judge Information about Students from Ethnic Minorities and with Low Socioeconomic Background, in *Social Psychology of Education,* 17(4):589-607.

Goe, L. (2007): *The Link Between Teacher Quality and Student Outcomes: A Research Synthesis*, National Comprehensive Center for Teacher Quality.

Gold, R.M., Reilly, A., Silberman, R., Lehr, R. (1971): Academic Achievement Declines Under Pass-Fail Grading, in *Journal of Experimental Education*, 39(3):17–21.

Goldhaber, D., Lavery, L. Theobald, R. (2015): Uneven Playing Field? Assessing the Teacher Quality Gap between Advantaged and Disadvantaged Students, in *Educational Researcher*, 44(5):293-307.

Goldin, C., Katz, L.F., Kuziemko, I. (2006): The Homecoming of American College Women: The Reversal of the College Gender Gap, in *Journal of Economic Perspectives,* 20(4):133–156.

Goldin, C., Rouse, C. (2000): Orchestrating Impartiality: The Impact of “Blind” Auditions on Female Musicians, in *American Economic Review,* 90(4):715–41.

Gosling, D.A. (1968): Standardized Ability Tests and Testing, in *Science*, 159(3817):851-855.

Grimaldi, E., Serpieri, R. (2012): The transformation of the Education State in Italy: A Critical Policy Historiography from 1944 to 2011, in *Italian Journal of Sociology of Education*, 10(1):146-180.

Grimaldi, E., Serpieri, R. (2012): The transformation of the Education State in Italy: A Critical Policy Historiography from 1944 to 2011, in *Italian Journal of Sociology of Education*, 10(1):146-180.

Guskey, T.R., Link, L.J. (2019): Exploring the Factors Teachers Consider in Determining Students’ Grades, in *Assessment in Education: Principles, Policy & Practice*, 26(3):303-320.

Hales, I.W., Bain, P.T, Rand, L.P. (1971): *An Investigation of Some Aspects of the Pass-Fail Grading System*, Annual Meeting of the American Educational Research Association, New York.

Hallinger, P., Heck, R.H. (1998): Exploring the Principal’s Contribution to School Effectiveness: 1980-1995, in *School Effectiveness and School Improvement*, 9:157-191.

Hanushek, E.A., Lavy, V., Hitomi, K. (2008): Do Students Care about School Quality? Determinants of Dropout Behavior in Developing Countries, in *Journal of Human Capital*, 2(1):69-105.

Hattie, J. (2012): *Visible Learning for Teachers: Maximizing Impact on Learning.* Routledge/Taylor & Francis Group.

Heckman, J.J., Kautz, T. (2014): Fostering and Measuring Skills: Interventions that Improve Character and Cognition, in Heckman, J.J., Humphries, J.E., Kautz, T. (Eds): *The Myth of Achievement Tests: The GED and the Role of Character in American Life,* Univ of Chicago Press, Chicago, 341–430.

Helland, H. (2007): How Does Social Background Affect the Grades and Grade Careers of Norwegian Economics Students?, in *British Journal of Sociology of Education*, 28(4):489-504.

Hernán, M.A. Robins, J.M. (2006): Instruments for Causal Inference: An Epidemiologist's Dream?, *Epidemiology*, 17(4):360-372.

Hinnerich, B.T., Höglin, E., Johannesson, M. (2011): Are Boys Discriminated in Swedish High Schools?, in *Economics of Education Review,* 30(4):682–690.

Hinnerich, B.T., Hoglin, E., Johannesson, M. (2015): Discrimination Against Students with Foreign Backgrounds: Evidence from Grading in Swedish Public High Schools, in *Education Economics*, 23(6):660–676.

Hinton, D.P., Higson, H. (2017): A Large-Scale Examination of the Effectiveness of Anonymous Marking in Reducing Group Performance Differences in Higher Education Assessment, in *PLoS ONE,* 12(8).

Hochweber, J., Hosenfeld, I., Klieme, E. (2013): Classroom Composition, Classroom Management, and the Relationship Between Student Attributes and Grades, in *Journal of Educational Psychology,* 106:1.

Hornstra, L., Stroet, K., van Eijden, E., Goudsblom, J., Roskamp, C. (2018): Teacher Expectation Effects on Need-Supportive Teaching, Student Motivation, and Engagement: A Self-determination Perspective, in *Educational Research and Evaluation,* 24(3–5):324–345.

Hoxby, C.M. (2002): The Power of Peers. How does the Makeup of a Classroom Influence Achievement?, in *Education Next*, educationnext.org

Iacus, S.M., Porro, G. (2008): Teachers' Evaluations and Students' Achievement: How to Identify Grading Standards and Measure their Effects, in *Education Economics*, 19(2):139-159.

IES (2009): NAEP Data Explorer, National Center for Education Statistics. http://nces.ed

INVALSI (2018): Rilevazione nazionale degli apprendimenti 2017-2018: Rapporto risultati. Roma: Istituto Nazionale per la Valutazione del Sistema Educativo di Istruzione e Formazione.

Isenberg, E., Max, J., Gleason, P., Potamites, L., Santillano, R., Hock, H., Hansen, M. (2013): *Access to Effective Teaching for Disadvantaged Students*, Washington, DC, National Center for Education Evaluation and Regional Assistance.

Isnawati, I., Saukah, A. (2017): Teachers' Grading Decision Making, in *TEFLIN Journal*, 28(2).

Jackson, M., Erikson, R., Goldthorpe, J.H., Yaish, M. (2007): Primary and Secondary Effects in Class Differentials in Educational Attainment: The Transition to A-Level Courses in England and Wales, in *Acta Sociologica,* 50(3);211–29.

Jackson, C.K. (2018): What Do Test Scores Miss? The Importance of Teacher Effects on Non–Test Score Outcomes, in *Journal of Political Economy*, 126(5):2072-2107.

Jackson, M. (2013): *Determined to Succeed? Performance versus Choice in Educational Attainment*, Stanford, Stanford University Press.

Jackson, M., Jonsson, J.O., Rudolphi, F. (2012): Ethnic Inequality in Choice-Driven Education Systems: A Longitudinal Study of Performance and Choice in England and Sweden, in *Sociology of Education*, 85(2):158-178.

Jæger, M.M. (2009): Equal Access but Unequal Outcomes: Cultural Capital and Educational Choice in a Meritocratic Society, in *Social Forces*, 87(4).

Jalava, N., Joensen, J.S., Pellas, E. (2015): Grades and Rank: Impacts of Non-financial Incentives on Test Performance, in *Journal of Economic Behavior & Organization*, 115:161-196.

Jussim, L. (1989): Teacher Expectations: Self-fulfilling Prophecies, Perceptual Biases, and Accuracy, in *Journal of Personality and Social Psychology*, 57(3):469-480.

Jussim, L., Eccles, J., Madon, S. (1996): Social Perception, Social Stereotypes, and Teacher Expectations: Accuracy and the Quest for the Powerful Self-fulfilling Prophecy Advances, in *Experimental Social Psychology,* 28:281-388.

Jussim, L., Harber, K.D. (2005): Teacher Expectations and Self-Fulfilling Prophecies: Knowns and Unknowns, Resolved and Unresolved Controversies, in *Personality and Social Psychology Review*, 9(2):131–155.

Kautz, T., Heckman, J., Diris, R., ter Weel, B., Borghans, L. (2014): *Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success*, n. w20749, National Bureau of Economic Research.

Khorramdel, L., Pokropek, A., Joo, S., Kirsch, I., Halderman, L. (2020): Examining gender DIF and gender differences in the PISA 2018 reading literacy scale: A partial invariance approach, in *Psychological Test and Assessment Modeling*, 62(2):179-231.

Kiss, D. (2013): Are Immigrants and Girls Graded Worse? Results of a Matching Approach, in *Education Economics,* 21(5):447–463.

Klapp Lekholm, A., Cliffordson, C. (2009): Effects of Student Characteristics on Grades in Compulsory School, in *Educational Research and Evaluation*, 15(1):1-23.

Kollmayer, M., Schober, B., Spiel, C. (2018): Gender Stereotypes in Education: Development, Consequences, And interventions, in *European Journal of Developmental Psychology, 15*(4):361–377.

Komarraju, M., Karau, S.J., Schmeck, R.R. (2009): Role of the Big Five Personality Traits in Predicting College Student's Academic Motivation and Achievement, in *Learning and Individual Differences,* 19:47-52.

Kunnath, J.P. (2017): Teacher Grading Decisions: Influences, Rationale, Practices, in *American Secondary Education*, 45(3).

Laidra, K., Pullmann, H., Allik, J. (2007): Personality and Intelligence as Predictors of Academic Achievement: A Cross-Sectional Study from Elementary to Secondary School, in *Personality and Individual Differences,* 1-11.

Lavy, V. (2008): Do Gender Stereotypes reduce Girls' or Boys' Human Capital Outcomes? Evidence from a Natural Experiment, in *Journal of Public Economics,* 92(10):2083–2105.

Lavy, V., Sand, E. (2015): On the Origins of Gender Human Capital Gaps: Short and Long Term Consequences of Teachers' Stereotypical Biases, in *NBER*, Working Paper Series.

Lechner, C., Danner, D., Rammstedt, B. (2017): How is Personality Related to Intelligence and Achievement? A Replication and Extension of Borghans et al. and Salkever, in *Personality and Individual Differences,* 111:86–91.

Levitt, S.D., List, J.A., Neckermann, S., Sadoff, S. (2012): The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance, in NBER Working Paper (18165).

Lewis, L., Parsad, B., Carey, N., Bartfai, N., Farris, E., Smerdon, B. (1999): Teacher Quality: A Report on the Preparation and Qualifications of Public School Teachers, in *Education Statistics Quarterly*, 1(1).

Lindahl, E. (2007): Comparing Teachers' Assessments and National Test Results: Evidence from Sweden, in Working Paper 24, IFAU – Institute for Evaluation of Labour Market and Education Policy.

Linden, A. (2014): MMWS: Stata Module to Perform Marginal Mean Weighting through Stratification, in *Statistical Software Components* S457886, Boston College Department of Economics.

Lipnevich, A.A., Guskey, T.R., Murano, D.M., Smith, J.K. (2020): What do Grades Mean? Variation in Grading Criteria in American College and University Courses, in *Assessment in Education: Principles, Policy & Practice,* 27(5):480–500.

Lord, F.M. (2012): *Applications of Item Response Theory to Practical Testing Problems*, New York: Routledge.

Loury, L.D., Garman, D. (1995): College Selectivity and Earnings, in *Journal of Labor Economics*, 13(2): 289–308.

Macaulay, R. (1978): The Myth of Female Superiority in Language, in *Journal of Child Language*, 5(2):353-363.

Machin, S., Pekkarinen, T. (2008): Global Sex Differences in Test Scores Variability, in *Science,* 322:1331–1332.

Machts, N., Kaiser, J., Schmidt, F.T.C., Möller, J. (2016): Accuracy of Teachers' Judgments of Students' Cognitive Abilities: A Meta-Analysis, in *Educational Research Review*, 19:85-103.

Manganelli, S., Cavicchiolo, E., Lucidi, F., Galli, F., Cozzolino, M., Chirico, A., Alivernini, F. (2021): Differences and Similarities in Adolescents' Academic Motivation Across Socioeconomic and Immigrant Backgrounds, in *Personality and Individual Differences*, 182:111077.

Marzano, R.J., Waters, T., McNulty, B. (2005): *School Leadership That Works: From Research to Results*, Aurora, CO: ASCD and McREL.

Mayer, S., Kalil, A., Oreopulos, P., Gallegos, S. (2015): Using Behavioural Insights to Increase Parental Engagement, *NBER* working paper 21602.

McCutcheon, A.L. (1987): *Latent Class Analysis*, SAGE Publications, Inc

McKown, C., Weinstein, R.S. (2002): Modeling the Role of Child Ethnicity and Gender in Children's Differential Response to Teacher Expectations, in *Journal of Applied Social Psychology*, 32:159−184.

McKown, C., Weinstein, R.S. (2003): The Development and Consequences of Stereotype-Consciousness in Middle Childhood, in *Child Development*, 74(2):498−515.

McMillan, J.H. (2001): Secondary Teachers' Classroom Assessment and Grading Practices, in *Educational Measurement: Issues and Practice*, 20(1):20–32.

McMillan, J.H. (2003). Understanding and Improving Teachers' Classroom Assessment Decision Making: Implications for Theory and Practice, in *Educational Measurement: Issues and Practice*, 22(4):34-43.

Mechtenberg, L. (2009): Cheap Talk in the Classroom: How Biased Grading at Schools Explains Gender Differences in Achievement, Career Choices and Wages, in *Review of Economic Studies*, 76:1431-1459.

Melissa, C., Sampo, C. Paunonen, V. (2007): Big Five Personality Predictors of Post-Secondary Academic Performance, in *Personality and Individual Differences*, 43:437-448.

Merton, R.K. (1957): Social Theory and Social Structure. Revised edition. N.Y.: The Free Press.

Mickelson, R.A. (1989): Why Does Jane Read and Write So Well? The Anomaly of Women's Achievement, in *Sociology of Education,* 62(1):47–63.

Miles, M.B., Huberman, A.M. (1994): *Qualitative Data Analysis: An Expanded Sourcebook.* Thousand Oaks, CA: Sage Publications.

Miur (2017): "*Le iscrizioni al primo anno delle scuole primarie, secondarie di primo e secondo grado del sistema educativo di istruzione e formazione.*" *Anno Scolastico* 2017–2018 Giugno 2017, *elaborazione su dati* MIUR – *Ufficio Statistica e Studi.*

Montanaro, P. (2008): I Divari Territoriali nella Preparazione degli Studenti Italiani: Evidenze dalle Indagini Nazionali e Internazionali." Questioni di Economia e Finanza. 14. Banca d'Italia.

Montmarquette, C., Mahseredjian, S. (1989): Could Teacher Grading Practices Account for the Unexplained Variation in School Achievements?, in *Economics of Education Review*, 8(4):335–343.

Morris, T.T., Davey Smith, G., van den Berg, G., Davies, N.M. (2021): Consistency of Noncognitive Skills and Their Relation to Educational Outcomes in a UK Cohort, in *Translational Psychiatry*,11(1):563.

Morris, T.T., Smith, G.D., van den Berg, G. Davies, N.M. (2021): Consistency of Noncognitive Skills and their Relation to Educational Outcomes in a UK Cohort, in *Translational Psychiatry*, 11(1):563.

Nguyen, H.T., Connelly, L.B. Le, H.T., Mitrou, F., Taylor, C., Zubrick, S.R. (2019): *Sources of Ethnicity Differences in Non-Cognitive Development in Children and Adolescents,* MPRA Paper No. 96785.

OECD (2009): Equally Prepared for Life? How 15-years-old Boys and Girls Perform in School. Paris: OECD (Organisation for Economic Co-Operation and Development).

OECD (2014): Are Boys and Girls Equally Prepared for Life? Paris: OECD (Organisation for Economic Co-Operation and Development).

OECD (2019): PISA 2018: Insights and Interpretation.

OECD (forthcoming): PISA 2018 Technical Report, OECD Publishing, Paris.

Ouazad, A. (2008): *Assessed by a Teacher Like Me: Race, Gender and Subjective Evaluations*, Centre for the Economics of Education DP 98.

Panichella, N., Triventi, M. (2014): Social Inequalities in the Choice of Secondary School, in *European Societies,* 16(5):666–693.

Pattison, E., Grodsky, E., Muller, C. (2013): Is the Sky Falling? Grade Inflation and the Signaling Power of Grades, in *Educational Researcher*, 42:259–265.

Pavolini, E., Argentin G., Barbieri G., Falzetti P., Ricci, R. (2015): L'Influenza delle Scuole e del Contesto Locale sui Divari Territoriali delle Competenze degli Studenti, in Asso F., Pavolini, E. (eds): *L'istruzione Difficile. I Divari nelle Competenze tra Nord e Sud*, Roma, Donzelli.

Pei, Z., Pischke, J.S., Schwandt, H. (2019): Poorly Measured Confounders are More Useful on the Left than on the Right, in *Journal of Business & Economic Statistics*, 37(2):205-216.

Pekkarinen, T. (2012): Gender Differences in Education, in *Nordic Economic Policy Review,* 1(1):165–94.

Pensiero, N., Giancola, O., Barone, C. (2019): Chapter 5: Socioeconomic Inequality and Student Outcomes in Italian Schools, in Volante, L., Schnepf, S., Jerrim, J., and Klinger, D. (Eds): *Socioeconomic Inequality and Student Outcomes: National Trends, Policies, and Practices*, Education Policy and Social Inequality Series, New York, Springer, 81-94.

PISA (2019): *PISA 2018 Results* (Volume II).

Pitts, D. (2007): Representative Bureaucracy, Ethnicity, and Public Schools, in *Administration and Society*, 39:497-526.

Popham, W.J. (1999): Why Standardized Tests Don't Measure Educational Quality, in *Educational Leadership-Using Standards and Assessment*, 56(6):8-15.

Protivínský, T., Münich, D. (2018): Gender Bias in Teachers' Grading: What is in the Grade, in *Studies in Educational Evaluation,* 59:141–149.

Puhani, P.A. (2018): Do Boys Benefit from Male Teachers in Elementary School? Evidence from Administrative Panel Data, in *Labour Economics*, 51:340-354.

Randall, J., Engelhard, G. (2010): Examining the Grading Practices of Teachers, in *Teaching and Teacher Education*, 26:1372-1380.

Raudenbush, S. W., Eschmann, R. D. (2015): Does Schooling Increase or Reduce Social Inequality, in *Annual Review of Sociology,* 41(1):443-470.

Ready, D.D, Wright, D.L. (2011): Accuracy and Inaccuracy in Teachers' Perceptions of Young Children's Cognitive Abilities: The Role of Child Background and Classroom Context, in *American Educational Research Journal*, 48:335-360

Reimer, D. (2019): Teachers and Teacher Education: A Call for a Renewed Sociological Research Agenda, in *International Studies in Sociology of Education,* 28(2):90–109.

Ridgeway, C.L., Smith-Lovin, L. (1999): The Gender System and Interaction, in *Annual Review of Sociology*, 25:191–216.

Riegle-Crumb, C., Humphries, M. (2012): Exploring Bias in Math Teachers' Perceptions of Students' Ability by Gender and Race/Ethnicity, in *Gender & Society,* 26(2):290–322.

Rindermann, H. (2007): The Relevance of Class Ability for Teaching and Development of Individual Competences, *Unterrichtswissenschaft,* 35:68–89.

Rist, R.C. (1977): On the Relations among Educational Research Paradigms: From Disdain to Détente, in *Anthropology and Education Quarterly*, 8(2):42-49.

Robinson, V.M.J., Lloyd, C.A., Rowe, K.J. (2008): The Impact of Leadership on Student Outcomes: An Analysis of the Differential Effects of Leadership Types, in *Educational Administration Quarterly*, 44(5):635-674.

Rosenthal, R., Jacobson, L. (1968): *Pygmalion in the Classroom: Teacher Expectations and Student Intellectual Development*, New York: Holt.

Rubie-Davies, C. M., Peterson, E. R., Sibley, C. G., Rosenthal, R. (2015): A Teacher Expectation Intervention: Modelling the Practices of High Expectation Teachers, in *Contemporary Educational Psychology,* 40:72–85.

Rubie-Davies, C.M. (2006): Teacher Expectations and Student Self-perceptions: Exploring Relationships, in *Psychology in the Schools*, 43(5):537-552.

Rubie-Davies, C.M. (2018): *Teacher Expectations in Education*, New York, NY: Routledge.

Sass, T.R., Hannaway, J., Xu, Z., Figlio, D.N., Feng, L. (2012): Value Added of Teachers in High-Poverty Schools and Lower Poverty Schools, in *Journal of Urban Economics*, 72(2):104-122.

Schneider, T. (2014): Reviwed Work: Determined to Succeed? Performance versus Choice in Educational Attainment, in *European Sociological Review*, 30(83):410-412.

Schofer, E., Meyer, J.W. (2005): The Worldwide Expansion of Higher Education in the Twentieth Century, in *American Sociological Review,* 70(6):898–920.

Shavit, Y., Blossfeld, H. (1993): *Persistent Inequality: Changing Educational Attainment in Thirteen Countries,* Social Inequality Series, Westview Press.

Speybroeck, S., Kuppens, S., Van Damme, J., Van Petegem, P., Lamote, C., Boonen, T., de Bilde, J. (2012): The Role of Teachers' Expectations in the Association Between Children's SES and Performance in Kindergarten: A Moderated Mediation Analysis, in *PloS one,* 7(4).

Spilt, J.L., Koomen, H.M.Y., Jak, S. (2012): Are Boys Better Off with Male and Girls with Female Teachers? A Multilevel Investigation of Measurement Invariance and Gender Match in Teacher–Student Relationship Quality, in *Journal of School Psychology*, 50(3):363-378.

Sprietsma, M. (2013): Discrimination in Grading: Experimental Evidence from Primary School Teachers, in *Empirical Economics*, 45(1):523–538.

Steel, C.M. (1997): A Threat in the Air: How Stereotypes Shape Intellectual Identity and Performance", in *American Psychologist*, 65(5):797-811.

Steele, C.M., Aronson, J. (1995): Stereotype Threat and the Intellectual Test Performance of African Americans, in *Journal of Personality and Social Psychology*, 69:797−811.

Stockè, V. (2007): Explaining Educational Decision and Effects of Families' Social Class Position: An Empirical Test of the Breen-Goldthorpe Model of Educational Attainment, in *European Sociological Review*, 23(4):505-519.

Strand, S. (2012): The White British–Black Caribbean Achievement Gap: Tests, Tiers and Teacher Expectations, in *British Educational Research Journal*, 38(1):75-101.

Stronge, J.H., Ward, T.J., Grant, L.W. (2011): What Makes Good Teachers Good? A Cross-Case Analysis of the Connection Between Teacher Effectiveness and Student Achievement, in *Journal of Teacher Education*, 62(4):339–355.

Tenenbaum, H.R., Ruck, M.D. (2007): Are Teachers' Expectations Different for Racial Minority than for European American Students? A Meta-analysis, in *Journal of Educational Psychology*, 99(2):253-273.

Teney, C., Devleeshouwer, P., Hanquinet, L. (2013): Educational Aspirations Among Ethnic Minority Youth in Brussels: Does the Perception of Ethnic Discrimination in the Labour Market Matter? A Mixed-method Approach, in *Ethnicities*, 13(5):584-606.

Terrier, C. (2015): Giving a Little Help to Girls? Evidence on Grade Discrimination and its Effect on Students' Achievement, *CEP*, Discussion Paper 1341.

Thorsen, C., Cliffordson, C. (2012): Teachers' Grade Assignment and the Predictive Validity of Criterion-Referenced Grades, in *Educational Research and Evaluation*, 18:153–172.

Thys, S. (2018): The Tertiary Effect of Social Class. Multilevel Studies on the Role of the Primary School (Teacher) in Educational Decision-making, PhD Thesis, Ghent University.

Timmermans, A.C., de Boer, H., van der Werf, M.P.C. (2016): An Investigation of the Relationship Between Teachers' Expectations and Teachers' Perceptions of Student Attributes, in *Social Psychology of Education*, 19(2):217-240.

Tobisch, A., Dresel, M. (2017): Negatively or Positively Biased? Dependencies of Teachers' Judgments and Expectations based on Students' Ethnic and Social Backgrounds, in *Social Psychology of Education: An International Journal,* 20(4):731–752.

Triventi, M. (2020): Are Children of Immigrants Graded Less Generously by their Teachers than Natives, and Why? Evidence from Student Population Data in Italy, in *International Migration Review* 54(3):765–795.

Tyner A., Gershenson S. (2020): Conceptualizing Grade Inflation, in *Economics of Education Review*, 78.

Urhahne, D. (2015): Teacher Behavior as a Mediator of the Relationship between Teacher Judgment and Students' Motivation and Emotion, in *Teaching and Teacher Education*, 45:73-82.

Urhahne, D., Wijnia, L. (2021): A Review on the Accuracy of Teacher Judgments, in *Educational Research Review*, 32:100374.

van den Bergh, L., Denessen, E., Hornstra, L., Voeten, M., Holland, R. W. (2010): The Implicit Prejudiced Attitudes of Teachers: Relations to Teacher Expectations and the Ethnic Achievement Gap, in *American Educational Research Journal*, 47(2):497-527.

van Ewijk, R. (2011): Same Work, Lower Grade? Student Ethnicity and Teachers' Subjective Assessments, in *Economics of Education Review*, 30:1045–1058.

Van Houtte, M., Demanet, J., Stevens, P.A.J. (2013): Curriculum Tracking and Teacher Evaluations of Individual Students: Selection, Adjustment or Labeling?, in *Social Psychology of Education*, 16(3):329-352.

Vecchione, M., Alessandri, G., Marsicano, G. (2014): Academic Motivation Predicts Educational Attainment: Does Gender Make a Difference?, in *Learning and Individual Differences*, 32:124-131

Vegas, E., De Laat., J. (2003): Do Differences in Teacher Contracts Affect Student Performance? Evidence from Togo, World Bank.

von der Embse, N., Jester, D., Roy, D., Post, J. (2018). Test Anxiety Effects, Predictors, and Correlates: A 30-year Meta-Analytic Review, in *Journal of Affective Disorders*, 227:483–493.

Walvoord, B., Anderson, V.J. (1998): *Effective Grading: A Tool for Learning and Assessment*, San Francisco, CA: Jossey-Bass

Weinstein, R.S., Middlestadt, S.E. (1979): Student Perceptions of Teacher Interactions with Male High and Low Achievers, in *Journal of Educational Psychology*, 71(4):421−431.

Wikström, M., Wikström, C. (2014): Who Benefits from University Admissions Tests?: A Comparison Between Grades and Test Scores as Selection Instruments to Higher Education, Vol 874 of Umeå economic studies, Department of Economics.

Williams, G.A. Kibowski, F. (2016): Latent Class Analysis and Latent Profile Analysis. In: *Handbook of Methodological Approaches to Community-Based Research: Qualitative, Quantitative, and Mixed Methods*. Oxford University Press, New York, 143-151.

Willingham, W.W., Pollack, J.M., Lewis, C. (2002): Grades and Test Scores: Accounting for Observed Differences, in *Journal of Educational Measurement*, 39(1):1-37.

Wilson, B.R. (1962): The Teacher's Role. A Sociological Analysis, in *The British Journal of Sociology*, 13(1):15-32.

Wilson, V. (2006): *Does Small Really Make a Difference? An Update*. A Review of the Literature on the Effects of Class Size on Teaching Practice and Pupils' Behaviour and Attainment, Scottish Council for Research in Education (SCRE) Centre: University of Glasgow.

Wright, S. P., Horn, S.P., Sanders, W.L. (1997): Teacher and Classroom Context Effect on Student Achievement. Implications for Teacher Evaluation, in *Journal of Personnel Evaluation in Education,* 11:57–67.

Zhu, M., Urhahne, D., Rubie-Davies, C.M. (2018): The Longitudinal Effects of Teacher Judgement and Different Teacher Treatment on Students' Academic Outcomes, in *Educational Psychology,* 38(5):648–668.