European Commission

JRC WORKING PAPERS AND PRE-PRINTS

Teacher Bias in Assessments by Student Ascribed Status:

A Factorial Experiment on Discrimination and Cultural Reproduction

JRC Working Papers Series on
Social Classes in the Digital Age
2024/02

Carlos J. Gil-Hernández
Irene Pañeda-Fernández
Leire Salazar
Jonatan Castaño Muñoz

Joint
Research
Centre

This Working Paper is part of a Working paper series on Social Classes in the Digital Age by the Joint Research Centre (JRC) The JRC is the European Commission's science and knowledge service. It aims to provide evidence-based scientific support to the European policymaking process. The scientific output expressed does not imply a policy position of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use that might be made of this publication. The Working paper series on Social Classes in the Digital Age is intended to give visibility to early-stage research to stimulate debate, incorporate feedback and engage into further developments of the research. This Working Paper is subject to the Commission Reuse Decision which allows authors to reuse the material without the need of an individual application.

Contact information
Name: Carlos J. Gil-Hernández
Address: Joint Research Centre, European Commission (Seville, Spain)
Email: carlos.GIL-HERNANDEZ@ec.europa.eu

EU Science Hub
https://joint-research-centre.ec.europa.eu/index_en

https://joint-research-centre.ec.europa.eu/knowledge-research/centre-advanced-studies/digclass_en

JRC136851

Seville: European Commission, 2024

How to cite this report: Gil-Hernández, C.J., Pañeda-Fernández, I., Salazar, L., Castaño Muñoz, J. *Teacher Bias in Assessments by Student Ascribed Status: A Factorial Experiment on Discrimination and Cultural Reproduction.* European Commission, Seville, 2024, JRC136851

Teacher Bias in Assessments by Student Ascribed Status: A Factorial Experiment on Discrimination and Cultural Reproduction

Carlos J. Gil-Hernández
(Joint Research Centre, European Commission, Seville)
Irene Pañeda-Fernández
(WZB Berlin Social Science Center)
Leire Salazar
(Joint Research Centre, European Commission, Seville)
Jonatan Castaño Muñoz
(University of Seville)

Abstract

Fair evaluations are fundamental for equal opportunity, with teachers as gatekeepers of academic merit in educational systems. Still, identifying their direct role in reproducing or mitigating inequalities via assessments is empirically challenging, yielding inconsistent findings on teacher bias from observational and experimental studies. We test interdisciplinary theories of status characteristics beliefs, statistical discrimination, and cultural reproduction with a pre-registered factorial experiment run on a large representative sample of Spanish pre-service teachers (n=1,717). This design causally identifies, net of true academic competence, the impact of student-ascribed status characteristics—gender, migrant and class origins—and cultural capital on teacher short- and long-term assessments, improving prior studies' limitations regarding theory testing, confounding, and power. Findings reveal teacher bias in an immediate task of essay grading favoring girls and highbrow cultural capital signals, aligning with status characteristics and cultural reproduction theories, respectively. Concerning teachers' long-term expectations, findings hint at statistical discrimination against boys, migrant-origin, and working-class students under uncertain information. Unexpectedly, ethnic discrimination changes from teachers favoring native origin in long-term expectations to migrant origin in essay evaluations, suggesting *compensatory* grading practices. These findings dig deeper into the complex roots of discrimination in teacher assessments as a mechanism underlying educational (in)equality.

Keywords: educational inequality, teacher bias, discrimination, assessments, sociology, social psychology, factorial survey experiment

Authors: Carlos J. Gil-Hernández (Joint Research Centre, European Commission, Seville), Irene Pañeda-Fernández (WZB Berlin Social Science Center), Leire Salazar (Joint Research Centre, European Commission, Seville), Jonatan Castaño Muñoz (University of Seville)

Acknowledgements: This project has been funded through the JRC Centre for Advanced Studies and the project Social Classes in the Digital Age (DIGCLASS). Jonatan Castaño Muñoz acknowledges the support of a) the 'Ramón y Cajal' grant RYC2020-030157 funded by MCIN/AEI/10.13039/501100011033 and by "ESF Investing in your future"; and b) University of Seville "VI University research plan" (VI plan propio de investigación). We are grateful to William Foley and Zbigniew Karpiński for their feedback to improve the quality of the paper.

Joint Research Centre reference number: JRC136851

Related publications and reports:

Experiment Pre-registration:

Gil-Hernández, C. J., Pañeda-Fernández, I., Salazar, L., & Castaño-Muñoz, J. (2023, March 31). Teacher's Bias in Assessments: A Factorial Survey Experiment. *Open Science Foundation*. https://doi.org/10.17605/OSF.IO/DZB3S

Experiment Dataset:

Carlos J. Gil Hernandez; Leire Salazar; Jonatan Castaño Muñoz; Irene Pañeda-Fernandez (2023): Teacher's Bias Dataset: A Factorial Survey Experiment. European Commission, Joint Research Centre (JRC) [Dataset] PID: http://data.europa.eu/89h/f14f5209-f032-4218-a89a-4643143809af

Passaretta, G., and Gil-Hernández, C.J. The Early Roots of the Digital Divide: Socioeconomic Inequality in Children's ICT Literacy from Primary to Secondary Schooling, European Commission, Seville, 2022, JRC128931.
https://joint-research-centre.ec.europa.eu/publications/early-roots-digital-divide-socioeconomic-inequality-childrens-ict-literacy-primary-secondary_en

Contents

1. Introduction

Students ascribed characteristics strongly shape inequality in educational outcomes (Breen and Jonsson 2005). Pupils from high socio-economic status (SES) (Chmielewski 2019), non-migrant backgrounds (Heath and Brinbaum 2007; Kao and Thompson 2003), and girls (DiPrete and Buchmann 2013; Mickelson 1989) systematically excel at school. The role of families (Jackson 2013) and school context (Passaretta and Skopek 2021; Downey and Condron 2016) has been extensively scrutinized to explain persistent achievement gaps by student-ascribed status (Skopek and Passaretta 2021). Teachers' attitudes and characteristics (Jennings and DiPrete 2010), however, received less attention despite documented disparities between their assigned grades and students scores in blindly assessed standardized tests (Meissel et al. 2017; Südkamp et al. 2012)—a residual approach interpreted as evidence of teacher grading bias.

Teachers are the primary evaluators and gatekeepers of academic merit in the educational system (Bourdieu and Passeron 1990), yet identifying their direct role in reproducing or mitigating educational inequalities via assessments is empirically challenging (Jæger 2022). Like any human being, teachers are susceptible to implicit and explicit biases in their information processing, beliefs, attitudes, and stereotypes about students' individuals and groups, potentially influencing their cognition and evaluations (Fazio et al. 2023; Lorenz et al. 2023; Alesina et al. 2018). Such assumptions may lead to self-fulfilling prophecies impeding student progress (Carlana 2019; Spinath and Spinath 2005) since grades are the main signals for students and families to make educational decisions (Holm, Hjorth-Trolle and Jæger 2019). Therefore, understanding how teachers form their assessment practices is crucial to fostering fair evaluations and equal opportunity in education.

Emerging behavioral (Carlana, La Ferrera and Pinotti 2022; Alesina et al. 2018) and experimental studies (Gilgen and Stocker 2022; Owens, 2022; Geven et al. 2021; Quinn 2020; Wenz and Hoenig 2020; Tobisch and Dresel 2017; Glock et al. 2015; Sprietsma 2013; Hanna and Linden 2012; Auwarter and Aruguete 2010) indeed document that teacher assessments might depend on student ascribed features. Likewise, previous observational research identified a residual effect of teacher bias in grading (Schuessler and Sønderskov 2023), expectations (Timmermans, Kuyper and Werf 2015), and tracking or grade retention recommendations (Batruch et al. 2023; Carlana, La Ferrera and Pinotti 2022; Salza 2022; Timmermans et al. 2018) as a function of students' ascribed characteristics[1], namely, gender (Marcenaro-Gutierrez, Prieto-Latorre and Sánchez Rodriguez 2023; Carlana 2019), ethnic origin (Kisfalusi, Janky and Takács 2021; Triventi 2019; Alesina et al. 2018; Botelho, Madeira and Rangel 2015),

[1] For the sake of simplicity, from now on we refer to cultural capital as an ascribed status characteristic in addition to gender, ethnic and socioeconomic background, but, as explained in *section 2.2.3*, cultural capital is not necessarily an ascribed factor.

SES (Gortázar, Martínez de Lafuente and Vega-Bayo 2022.), and cultural capital (Jæger 2022; Jæger and Møllegaard 2017).

Despite accumulating evidence, teacher biases in assessments remain poorly understood due to omitted variable bias (i.e., unobserved socio-emotional skills) and measurement error in test scores with observational data (van Huizen, Jacobs and Oosterveen 2024), weak reliability of cognitive measures of direct implicit bias (Miles, Charron-Chénier and Cyrus Schleifer 2019), as well as insufficient statistical power (Schuessler and Freitag 2020) and external validity (Krolak-Schwerdt et al. 2017) in experimental studies (Petzold 2022). Furthermore, most previous research only tested single discrimination theories (Correll and Benard 2006) and focused on ethnic or gender discrimination (Zanga and De Gioannis 2023), leaving SES- and cultural capital-based biases under-researched while not disentangling the causal effect of all these students' ascribed characteristics on teacher assessments (Wenz and Hoenig 2020).

In this article, we contribute by testing if teachers show biases in assessments due to several students' ascribed status factors and framing our pre-registered hypotheses[2] in an interdisciplinary theoretical framework of discrimination spanning sociology, psychology, and economics. We draw on theories of status characteristics beliefs (Ridgeway 2014; Foschi 2000), implicit bias (Greenwald and Banaji 1995), statistical discrimination (Arrow 1998; 1973), and cultural reproduction (Jæger and Breen 2016; Lamont et al. 2014), which might operate simultaneously, to hypothesize negative bias in teachers' assessments of students with the following characteristics: boys, ethnic-minority origin, working-class background, and lowbrow cultural capital. We aim to partially tease out these theories' explanatory power by comparing three different educational outcomes from primary to secondary education, which convey different degrees of information and uncertainty for teacher evaluations.

Our pre-registered experimental and sampling design further contributes to studying the role of teacher biases and discrimination in educational inequality on four main methodological fronts. First, we study our research questions in a pre-registered experimental setting. Individual biases and stereotypes according to students' backgrounds are hard to capture with standard observational data due to social desirability bias and the impossibility of measuring all (un)observable student characteristics—such as behavior (Ferman and Fontes 2023) and true ability (van Huizen, Jacobs and Oosterveen 2024), generally proxied with low-stakes competence tests (Südkamp, Kaiser, and Möller 2012). To address these issues, we designed a full factorial experiment (Hainmueller, Hopkins and Yamamoto 2014) with 128 students' profiles—2^7 levels and dimensions—to isolate the causal effect of students' ascribed characteristics on teacher's assessments—essay grading, grade retention

[2] In the pre-registered pre-analysis plan, we did not specify different hypotheses for the empirical expectations expressed here (see below) as hypotheses 1, 2 and 3 on implicit bias or status characteristics theory (H1), statistical discrimination (H2), and cultural capital reproduction theories (H3), respectively. We formalized them jointly in the pre-analysis plan, but the same predictions by student's ascribed factors hold here.

recommendations, and expectations about enrolment in the upper-secondary academic track. To correctly identify the net effect of a student's ascribed characteristics—parental class, ethnic background (Spanish or Moroccan origin), gender—and cultural capital, we control for three additional ability dimensions to avoid confounding. These are students' language skills (objective essay quality), the number of failed/passed subjects, and socio-emotional skills (classroom behavior and effort) in the current academic term. Our study took place online, where we randomly allocated each participant one fictitious profile of an elementary education student in 6th grade to evaluate.

Second, laboratory and factorial survey experiments are often criticized for their generally lower validity (Petzold, 2022; Krolak-Schwerdt et al. 2017) relative to field experiments or behavioral tests directly measuring automatic cognition, such as the *Implicit Association Test* (IAT) (Melamed et al. 2019; Miles, Charron-Chénier and Schleifer 2019), even though these tests are subject to validity issues of their own (Mitchell and Tetlock 2017; Arkes and Tetlock 2004). Thus, to increase the validity of our factorial design and the signaling power of the survey instruments, we randomly assigned different versions of a real task (Wenz and Hoenig 2020), an essay written by a sixth-grade student, experimentally manipulating its cultural capital signals (low vs high) and objective quality—previously pre-tested with 243 in-service elementary education teachers. Furthermore, thanks to the full factorial design, we untangle parental social class from ethnic origin. In previous experimental research, students' SES was subtly signaled through names and surnames (Wenz and Hoenig 2020). However, as participants might not correctly identify SES variation within foreign-origin names (Crabtree et al. 2022), we embed the students' SES signal (i.e., father's occupation) within a fictitious student file, resembling the real ones used in schools, including academic records, and within the essay. In the file, we signal the student's gender and ethnic origin with name and surname, as well as family SES through the family contact email. Thus, a significant contribution of our study is to disentangle, to our knowledge for the first time experimentally, cultural capital signals (Breinholt and Jæger 2019) from a student's objective ability (Farkas 2003), parental social class and ethnic origin using realistic and externally validated instruments.

Third, instead of in-service teachers, we sampled pre-service teachers—students enrolled in the BA in Primary Education—to identify if they already show biases in assessments well before interacting with actual students or being exposed to the institutional school context. Throughout their careers, teachers might sort into schools with socio-demographic characteristics and organizational processes aligned with their previous biases and ascribed traits (Lievore and Triventi 2023). At the same time, school-level institutional factors and classroom composition (Schuessler and Sønderskov 2023) might reinforce or mitigate preexisting teacher biases (Pit-ten Cate and Glock 2019). Thus, focusing on pre-service teachers might establish a benchmark for *inter-group relations* studies (Elwert, Keller and Kotsadam 2023) while informing

the debate on early interventions to promote fairness and antidiscrimination in teacher training programmes (Lehmann-Grube, Tobisch and Dresel 2023; Alesina et al. 2018).

Fourth, experimental surveys usually have lower statistical power and representativeness than large-scale surveys. Thus, we went beyond a small convenience sample, implementing a systematic random sampling with probability proportional to size (OECD 2020). We recruited a representative sample of 19 public and private Spanish universities and contacted all students enrolled in the *Primary Education Teaching BA* to reach 1,717 valid respondents, which allows for identifying powered main effects according to a pre-registered power plan.

Findings indicate teacher biases in essay grading, showing a preference for girls and students signaling high cultural capital. These results align with theories related to status characteristics, implicit bias, and cultural reproduction. In terms of teachers' future educational expectations, the results suggest a form of statistical discrimination against boys, students with migrant origins, and working-class backgrounds. Surprisingly, the ethnic bias shifts from favoring native-origin students in teachers' long-term expectations to *compensatory* grading in essay assessments favoring those with a migrant background. We delve into the theoretical implications of these findings on the intricate origins of teachers' discriminatory tendencies and biased assessments that contribute to educational inequalities.

This article is organized as follows. First, we review the main theories and previous findings accounting for teacher bias and the mechanisms at work to formulate our research hypotheses. Second, we explain our sampling procedure and the experimental research design, describing variable operationalization and model specifications. Third, we describe the empirical results. Finally, we discuss the implications of our findings for the interdisciplinary literature on educational inequality and discrimination and conclude by highlighting limitations that pave the way for future research.

2. Theoretical background, previous findings and hypotheses

In this section, we focus on the theories that can explain how teachers, as institutional gatekeepers, contribute to creating observed achievement gaps by student-ascribed characteristics. Below, we expand on each theory and how to differentiate between them by focusing on their observable, testable implications while reviewing related previous findings.

2.1. Implicit bias and status characteristics beliefs

Theories of implicit biases focus on how micro-processes contribute to creating social inequality in subtle ways (Fazio et al. 2023; Greenwald and Banaji 1995). Implicit cognition (Greenwald and Krieger 2006) is an automatic process that happens outside of one's conscious attentional focus. Implicit bias theories can account for discrimination via two interrelated processes of implicit cognition: (1) a tendency to like

or dislike members of a group (implicit attitudes), and (2) the association of a group with a particular positive or negative trait (implicit stereotypes). Accordingly, studies deploying implicit bias tests in the educational context broadly identified teachers' negative reactions against immigrants and low-SES students (Carlana, La Ferrera and Pinotti 2022; Pit-ten Cate and Glock 2019; Alesina et al. 2018), while results on gender are more mixed (Carlana 2019; Glock and Klapproth 2017). These implicit associations do not necessarily align with explicit attitudes, stereotypes or judgment behavior (Glock and Sabine Krolak-Schwerdt 2014), and might remain unconscious until triggered.

The psychological theory explaining implicit biases can also be understood through a sociological lens if we consider that implicit biases emerge during early socialization and are stored in implicit memory as cultural schemata (DiMaggio 1997). These, in turn, give rise to biases because people are more likely to perceive and recall information consistent with preexisting mental structures (DiMaggio 1997).

Sociologists have also developed separate theories with similar implications to the implicit bias theory (Melamed et al. 2019). *Status Characteristics Theory* (SCT) aims to explain social inequality by focusing on *status* distinctions—*beliefs* about which social groups are more competent or deserving (Ridgeway 2014; Berger et al. 1977). The crux of the argument is that such beliefs naturally emerge in small-group interactions, and they will fall along categorical stratification groups—ethnicity, gender, and social class (Foley 2023)—as long as these ascribed characteristics convey distinctive status values and are visible and salient to the task (Ridgeway 2014). In sum, SCT is a theory of *status generalization* that attributes specific abilities to individuals based on their status characteristics (Correll and Ridgeway 2006:33). Individuals may not fully realize that they hold differential *expectations* of competence for people depending on their ascribed characteristics, linking this theoretical perspective developed by sociologists to psychological theories of implicit bias (Melamed et al. 2019).

Similar status generalization processes might arise in the educational context where teachers make comparative performance evaluations of students (Kisfalusi, Janky and Takács 2018). Thus, the theories discussed above imply that part of the observed gaps in academic performance by ascribed characteristics (namely, gender, ethnicity, and SES) might reflect teachers' implicit beliefs and expectations about competence and deservingness. These differential expectations by students' ascribed characteristics alone can generate gaps via self-fulfilling prophecies (Merton 1968). Crucially, as the *Double Standards Theory* posits (Foschi 2000), a branch of the SCT, *standards* tied to status characteristics might result in differential performance expectations and biased ability assessments among equally competent students. Due to entrenched status beliefs, lower-status individuals must outperform higher-status peers for equal task competence recognition since high performance would be inconsistent with their bottom-status position. Hence, Double Standards Theory reveals harsher scrutiny for lower-status individuals, favoring lenient judgment for equally competent higher-status counterparts. Accordingly, teachers might contribute to reproducing the observed

educational gaps in favor of girls over boys, high SES over low SES students, and native over migrant-origin students, net of their objective academic ability.

To explain the gender bias, we argue that teachers might form their implicit biases and status characteristics beliefs by internalizing stereotypes about girls doing better than boys in school since, currently, girls objectively outperform boys in educational performance and attainment (DiPrete and Buchmann 2013), and girls are accordingly perceived as more academically competent (Homuth, Thielemann and Wenz 2023), particularly in language proficiency (Krkovic et al. 2014). The internalization process may begin during their schooling and teacher training years and be further reinforced as they begin teaching pupils. Further, we argue that this implicit bias may also be boosted via in-group bias (Kisfalusi, Janky and Takács 2018; Tajfel and Turner 1986), given that primary education teachers are overwhelmingly female. To explain ethnic and SES biases, we similarly argue that teachers may begin to be exposed to and internalize negative stereotypes about low-SES and migrant-origin individuals from an early age, given that these abound in the Western context in general and the Spanish case in particular (Cea D'Ancona 2016). During their previous role as students, teachers' status beliefs may be further reinforced during their schooling by exposure to these ascribed status groups as classmates, which actually underperform native-origin and high-SES students (Skopek and Passaretta 2021) and are perceived as less academically competent (Homuth, Thielemann and Wenz 2023).

In sum, when teachers perceive a student's ascribed status characteristic, their evaluation is tainted by their tendency to like or dislike particular groups or their differing expectations, beliefs and standards about the competence of individuals belonging to that group. In contrast with the *statistical discrimination* perspectives discussed below, the prediction is that teacher biases are cognitive in nature, remaining relatively stable and difficult to change with new information about a specific student. Instead, when teachers receive new input, they try to reconcile it with pre-existing status beliefs. Still, even though a single individual interaction is unlikely to change behavior, teachers embedded in a school consistently exposed to counter-stereotypical interactions could theoretically decrease their bias (Elwert, Keller and Kotsadam 2023).

Based on these considerations, if implicit cognitive biases and status characteristics beliefs are at play, we predict the following in *hypothesis 1 (H1):*

H1. Implicit bias, beliefs and standards about student status characteristics drive teacher evaluations by over-grading, recommending less grade retention, and expressing higher expectations for girls (vs boys), natives (vs migrant origin), and high-SES students (vs low-SES), independently of other correlated competence factors, like objective academic performance and socio-emotional skills, and cultural capital.

2.2. Statistical discrimination

Theories of statistical discrimination, mainly by economists (Arrow 1998, 1973; Borjas and Goldberg 1978; Aigner and Cain 1977; Phelps 1972), have a crucial distinction with

implicit bias or status characteristics theory. Rather than resulting from deep-rooted beliefs and expectations, discrimination happens due to a lack of perfect information and diminishes once obtained. The key idea in the original formulation applied to the labor market is that, under imperfect information about true employees' productivity or performance, the rational action to follow by an employer is to proxy unknown individual productivity using the employee's observable characteristics, such as gender or ethnicity. The information employers use from ascribed characteristics is the average performance of employees belonging to a given ascribed group, known from previous experience or historical knowledge. When given additional information to make an assessment, the prediction is that discrimination diminishes or even disappears.

Even though the theory was initially developed to explain hiring discrimination, recently, it has been applied to studying discrimination in the educational context. For instance, Hanna and Linden (2012) find experimental evidence of statistical discrimination in grading. When asked to evaluate a series of exams with randomly assigned ascribed characteristics (gender, age, and caste), they find that teachers' bias against low-caste students decreases as the evaluation process advances. The authors interpret this as evidence for statistical discrimination because they argue that evaluators rely less on demographic characteristics as they obtain information about the testing instrument and the grade distribution.

Likewise, Botelho and Rangel (2015) also interpret evidence of grading discrimination through the lens of statistical discrimination theory. The authors compare teacher assessments of 8th graders across 10.6 thousand public school classrooms in Brazil to the scores obtained in end-of-year standardized (blindly marked) proficiency tests to study racial discrimination. They use the length of classroom interaction time between the teacher and a given student as a proxy for the level of information that a teacher has on the student. They show no racial discrimination for those students graded by a teacher who had already taught them before 8th grade. Hence, racial discrimination in grading is only present for those attending classes with a teacher for the first time.

Studies on the impact of rubrics on assessment also uncover patterns of discrimination that could be compatible with the predictions of statistical discrimination theory: teachers' racial bias in grading is present with vague rubrics but disappears when using a rubric with more clearly defined evaluation criteria (Quinn 2020). Thus, teachers might rely less on students' ascribed characteristics as proxies for average performance under clear guidance on absolute evaluation (Hjorth-Trolle, Rosenqvist and Hed 2022).

In sum, a key implication differentiating the statistical discrimination theory from those discussed above is the expectation that the more information provided, the less discrimination. Applied to the context of explaining teachers' biases in assessment, we argue that statistical discrimination is less likely to be at play in specific task grading situations or end-of-year retention recommendations, where teachers can additionally

rely on rich information on a student's concurrent academic records and classroom behavior. One should note, though, that we do not consider implicit bias, SCT, or statistical discrimination mutually exclusive but complementary. All of them can be simultaneously at play while having different weights or explanatory power (Correll and Benard 2006). By contrast, we expect statistical discrimination to be more likely when teachers express long-term future educational expectations for individual students. Teachers often lack crucial information about the student's family life and other factors conditioning their future trajectory to make an informed prediction. In such a case, it seems more reasonable that they may use ascribed characteristics as a proxy.

If statistical discrimination is at play, we predict the following in *hypothesis 2 (H2)*:

H2. Under imperfect individual-level information, teachers express higher educational expectations for girls (vs boys), natives (vs migrant origin), and high-SES students (vs low-SES), net of other correlated competence factors, such as objective academic performance and socio-emotional skills, and cultural capital. Discrimination is generally larger and more likely to be explained by statistical discrimination than implicit bias or status characteristics beliefs when teachers express long-term expectations rather than grade a concrete task or recommend a short-term outcome (grade retention) under concurrent, detailed student information.

2.3. Cultural reproduction via cultural capital

Despite the popularity of classical theories of *cultural reproduction* via cultural capital, recent literature surveys conclude that we do not know much about the role of cultural capital in shaping social inequality (Jæger 2022). This article seeks to understand better the extent to which, in addition to the theories reviewed above, theories of cultural reproduction can explain teacher discrimination.

Such theories began to gain traction following Bourdieu's influential theory of cultural capital and cultural reproduction (Bourdieu and Passeron 1990; Bourdieu 1984; 1977; Bernstein 1961). Bourdieu defined cultural capital as high-status cultural signals that enhance social inequality (Lamont and Lareau 1988; Bourdieu 1984). Despite its early influence, cultural sociologists and stratification scholars have come to agree that core concepts and mechanisms are not well formalized in Bourdieu's original writings on cultural reproduction (van de Werfhorst 2010; Goldthorpe 2007; Kingston 2001; Lamont and Lareau 1988). Indeed, according to Jæger and Breen (2016:1108), "Research has yet […] to identify the specific mechanisms through which cultural capital may lead to educational success." Despite its shortcomings, cultural reproduction theories remain popular among education sociologists (Xu and Hampden-Thompson 2012; Yamamoto and Brinton 2010; van de Werfhorst and Hofstede 2007; Sullivan 2001; Roscigno and Ainsworth-Darnell 1999; Aschaffenburg and Maas 1997; DiMaggio 1982), demonstrating that the cultural capital concept can be a powerful instrument to identify educational inequality mechanisms if precisely formalized and tested (Jæger 2022).

Over the years, two different approaches have emerged when linking cultural capital to social stratification in education. The first approach follows from Bourdieu's proposition that cultural capital shapes gradients in students' educational attainment mainly via teachers' bias (Bourdieu 1984). Here, the proposition is that teachers interpret cultural capital as signals of academic brilliance *independently* of actual academic performance or ability. The argument operates through shared cultural scripts such as 'frames' or 'narratives' (Lamont et al. 2014; Small et al. 2010; Lamont and Small 2008). In other words, teachers are embedded in narratives that equate signals that fall high on the cultural capital gradient (Jæger et al. 2023) with academic brilliance, regardless of objective student performance. They learn to use and recognize such signals to gatekeep access to educational advancement, resulting in teacher discrimination in so far as they favor students who show these signals irrespective of their objective academic performance (Jæger and Breen 2016; DiMaggio 1982). In this way, teachers positively evaluate those children socialized in the dominant culture of the upper classes, to which most teachers belong and the school system legitimizes through canonical curricula (Bourdieu and Passeron 1990).

The second approach to studying how cultural capital shapes educational inequality departs from the classic Bourdieusian proposition that teachers are biased to reward cultural capital signals net of academic performance. Instead, the proposition is to conceive cultural capital as a set of socio-emotional or noncognitive skills (Jæger 2022) that are defined as "patterns of thought, feeling and behavior" (Almlund et al. 2011; Borghans et al. 2008: 974). In turn, these skills have been shown to nurture academic skills that directly improve children's educational success (Kisida et al. 2014; Kaufman and Gabler 2004) or to lead to an improved capacity to command attention and negotiate advantages in the classroom (Calarco 2014; Lareau 2011). As Jæger (2011:282) highlights, those children (and parents) who display high levels of cultural capital are also very likely to be the ones with high ability and motivation. That might lead to a considerable overestimation of the total effect of cultural capital.

Distinguishing between the two ways mentioned above of understanding the role of cultural capital is often not possible due to data limitations, and it remains unclear if there is a causal relationship between cultural capital and educational outcomes (Jæger 2022). This article thus aims to disentangle these two perspectives linking cultural capital to student educational success by testing for direct evidence of teacher bias as framed in the first perspective. We test whether teachers use performance-irrelevant cultural capital markers in their assessments, reinforcing categorical inequality over and above objective students' academic abilities and socio-emotional skills.

A key distinction with the other discrimination theories discussed above is that here, it would be signals of cultural capital, and not ascribed characteristics per se, that drive teacher discrimination. The implication is that teachers would respond to the cultural capital signals *regardless* of the ascribed characteristics of the student. Even if high-SES and non-migrant background students are more likely to show such signals in the

real world—all three factors orthogonal by design in our experiment, the claim is that a portrayal of cultural capital signals drives teachers' discrimination in their favor. In other words, it is an understanding of teachers' gatekeeping driven by a cultural signal rather than an ascribed characteristic.

If cultural reproduction is at play, we predict the following in *hypothesis 3 (H3)*:

H3. Teachers misconceive academic brilliance with highbrow cultural capital by over-grading, recommending less grade retention and expressing higher expectations for students signaling high cultural capital (vs low cultural capital), independently of other correlated ability and ascribed factors, such as objective academic performance and socio-emotional skills, parental SES, migrant background, and gender.

3. Data, methods, and variables

The hypotheses and research design were pre-registered on the *Open Science Foundation*[4] before data collection and analysis, and all data[5] and replication files are available on *GitHub*. Deviations from the pre-analysis plan are indicated where relevant.

3.1. Target population: Pre-service teachers

Our population of interest comprises students enrolled in any grade of the BA Degree in Primary Education in Spain. Holding this degree is a legal requisite to work as a teacher in public elementary schools. According to Spanish administrative data, in 2019/2020, only 9.8% of the students enrolled in the first grade of this BA Degree dropped out, and 3.2% enrolled in another degree (INE 2023). Furthermore, according to a Spanish survey with a representative sample of the graduates in Primary Education (INE 2020), in 2019, 5 years after graduation, 82% of the Primary Education graduates were employed, of those 76% of those as teachers (ISCO 22-23), 12% unemployed and 6% inactive (70% studying for the teachers' entry exam). Thus, most of our experimental sample of college students will eventually become teachers. Previous research also shows that pre-and in-service teachers are comparable regarding their biases magnitude, and the school context or teachers' characteristics do not seem to moderate them (Schuessler and Sønderskov 2023; Pit-ten Cate and Glock 2019). Likewise, Starck et al. (2020) showed that teachers and the general non-teacher American population display equivalent explicit racial and pro-White biases.

3.2. Sampling design and data

Experimental surveys, field, and lab experiments generally collect convenience samples that are not representative of the reference population and suffer from low statistical power and poor external validity. To address these issues, we drew a large-scale

[4] Gil-Hernández, C. J., Pañeda-Fernández, I., Salazar, L., & Castaño-Muñoz, J. (2023, March 31). Teacher's Bias in Assessments: A Factorial Survey Experiment. https://doi.org/10.17605/OSF.IO/DZB3S

[5] Carlos J. Gil Hernandez; Leire Salazar; Jonatan Castaño Muñoz; Irene Pañeda-Fernandez (2023): Teacher's Bias Dataset: A Factorial Survey Experiment. European Commission, Joint Research Centre (JRC) [Dataset] PID: http://data.europa.eu/89h/f14f5209-f032-4218-a89a-4643143809af

representative sample of 19 public and private Spanish faculties of education, reaching an analytical sample of 1,717 pre-service teachers—students enrolled in any grade of the 4-year BA Degree (or Double Degree) in Primary Education in the 2022/2023 academic year. To prevent obtaining a sample of only small or large faculties, we implemented an explicitly stratified systematic random sampling by public and private institutions with probability proportional to size (PPS) (see OECD 2020). We randomly (PPS) drew 20 institutions—15 public and 5 private to reflect the actual share of students in each explicit sampling stratum—from the population frame to be representative of all faculties of education across non-bilingual[6] Spanish regions (N=85 in 2020-22). We replaced 4 out of the 5 initially selected faculties with the next closest unit in the sampling frame according to the measure of size (enrolled students) due to non-response or refusal to participate. In total, 15 public and 4 private institutions participated in the study, and we invited 27,015 students (19,204 in public and 7,811 in private institutions; see online supplement A.2. and Table A.2.) to participate in the study via email from the faculty's dean or secretary. Participation was incentivized with a lottery of gift cards (see online supplement A.1.). Among all the 27,015 students enrolled in the BA Degree in Primary Education with emails listed in the faculty's directories, 1,028 students in 15 public (5.8% mean response rate) and 720 in 4 private faculties (11.5%) completed the online survey during the fieldwork between April and June 2023. We collected 1,748 observations (overall 7% response rate), which were finally reduced to 1,717 after excluding fraudulent or underage (age < 18) cases.[7]

Even though there are higher average response rates in private institutions, the overall share of students in these is virtually the same in the experimental sample (40%) compared to the actual population share (39.4%). As can be seen in Table 1, the socio-demographic characteristics of our experimental sample are generally representative of the population, even though there is a slight overrepresentation of females (+9.9%), foreign-born[8] (+3.4%) and older students (+9.8%), and an underrepresentation of students coming from highly-educated families (-10.7%) relative to administrative data on the whole reference population (INE, 2023). As indicated in the pre-analysis plan, to preserve the successful randomization, frequency, and distribution balance of vignettes over respondents, we did not apply any weights to adjust for participant characteristics within sampled universities in the main analyses (Schonlau et al. 2009). Nevertheless, as a robustness check, we generated and adjusted for calibration weights using raking estimators (Valliant and Dever 2018)—fed with the fully comparable population shares of the main socio-demographic variables (gender, age, and parental education)—to successfully replicate the findings from the main unweighted models (see online supplement section A.9. Table A.8.).

[6] Preventing regional identity and discrimination to confound ascribed characteristics (Polavieja 2023). We also excluded bilingual regions as our task involves Spanish competencies, which might vary by (non)bilingual regions.
[7] We ensured that participants provide honest and accurate responses by running attention checks (to drop those observations who replied too fast or completing the survey randomly) and identifying and filtering out duplicates.
[8] Note that this figure is not directly comparable as, in the administrative data, migrant-origin students are defined as non-Spanish nationality, while in our experiment we ask for the parental country of birth.

Table 1. *Sample and population characteristics*

	Population Data[a] 2022/2023	Experiment Sample 2023
Students (Institutions)	N=59,084 (94)	n=1,717 (19)
Total		
Students in Private Institutions	39.4%	40.0%
Female	68.8%	78.7%
Grade[d]		2.8
Age		
18-25	73.3%	63.4%
≥ 26	26.7%	36.6%
Foreign-Born Students	1.3%[b]	4.7%
Foreign-Born Parents		9.4%
Parental College Education	50.9%[c]	40.2%
Public Universities		
Students (Institutions)	N=35,785 (49)	n=1,030 (15)
Female	65.9%	77.2%
Grade[d]		2.7
Age		
18-25	90.3%	87.4%
≥ 26	9.7%	12.6%
Foreign-Born Students	1.4%[b]	3.3%
Foreign-Born Parents		9.0%
Parental College Education	50.2%[c]	41.4%
Private Universities		
Students (Institutions)	N=23,299 (45)	n=687 (4)
Female	73.1%	81.1%
Grade[d]		2.9
Age		
18-25	47.3%	27.4%
≥ 26	52.7%	72.6%
Foreign-Born Students	1.3%[b]	6.8%
Foreign-Born Parents		9.8%
Parental College Education	52.4%[c]	38.4%

Notes: (a) Administrative data (provisional) from the academic year 2022/2023, excluding non-bilingual regions. (b) Non-Spanish nationality. (c) Data from 2019/2020. (d) The average course of enrolment in the BA Degree in Primary Education, with SD=1.2 and ranging from 1 to 4 for the standard BA and from 1 to 5 for Double Degrees.

The pre-registered power analysis (see section A.3. of the online supplement) shows that the analytical sample exceeds the estimated 1,398 observations necessary to detect powered (80%) statistically significant coefficients according to the anticipated effect sizes at Cohen's D = 0.15. However, given the larger analytical sample drawn (n=1,717) but the smaller average effect sizes found (Cohen's D ≈ 0.1) in the experiment, we (re)estimated the minimum detectable effect sizes with power=0.8, two-sided alpha=0.05 and the observed SD of our three outcome variables at β = 0.133 (σ = 1.97) for essay grading, β = 0.199 (σ = 2.95) for grade retention recommendations and β = 0.146 (σ = 2.16) for academic track expectations (see online supplement A.3.). Most estimated coefficients lie above these powered thresholds.

3.3. Methods

3.3.1. Experimental design

We designed a full factorial experiment with 2^7 = 128 profiles or vignettes—seven dimensions and two levels (see Table 2 for an overview). As shown in Table 2 below, we experimentally manipulate student-ascribed status—gender, ethnicity, parental social class, and cultural capital. Besides these factors, to avoid omitted variable bias, we include three additional dimensions accounting for student's academic ability and behavior: student language-related skills signaled with an essay whose objective quality is experimentally controlled (see section 3.3.2. below), number of subjects failed, and socio-emotional skills (behavior and effort) in the current term evaluation (Ferman and Fontes 2023). In section 3.3.2., we explain in detail how we signal and operationalize each factor.

The vignette universe, or population, consists of $N_u = 2^7 = 128$ profiles or vignettes that are orthogonal by design. We implement a full factorial design, including all possible combinations or interactions to minimize standard errors and maximize estimation precision. The full factorial design allows identifying all main effects independently of each other and all two-way interactive terms (resolution V) exploiting the maximum variance (Auspurg and Hinz 2015). Even extreme combinations are plausible in the Spanish context. Thus, potentially non-realistic or implausible combinations in empirical terms are not excluded to avoid loss of efficiency.

Given the median response time identified in the experiment's pre-test (about 7.3 minutes), only one vignette or task was assigned to each respondent, known as a between-subject design, to avoid cognitive overload or fatigue (i.e., essay grading takes on average about 2.4 minutes), learning and response heuristics, and measurement error, so maximizing response rates and estimation precision. The vignettes are the analysis unit and are randomly assigned to respondents. In order to avoid confounding the vignette or experimental conditions with the respondent's characteristics, each vignette was rated by 14 different respondents, on average (ranging from 4 to 24), and we control for respondent-level covariates to increase the precision of the estimates (see section 3.3.3. below). We also ran collinearity tests among factors with measures

such as the variance inflation factor (VIF) and tolerance and a correlation matrix that yielded null results, thus indicating successful randomization (see online supplement section A.7. for a visual inspection). The online experiment and randomization of experimental conditions were implemented with *Qualtrics*® software. See online supplement A.2. for the structure and screens composing the online questionnaire.

Table 2. *Factors, levels and signalling*

Vignette Factors	Vignette Levels	Signaling
1. Gender	1. Female 0. Male	Directly and indirectly: (1) Student's gender and name in student's file; (2) and in essay's screen instructions.
2. Migrant background	1. Spanish origin (native majority) 0. Moroccan origin (largest ethnic minority)	Indirectly: (1) Student's name/surname in the student's file; (2) and in the essay's screen instructions; (3) Father's email (name and surname) in the student's file.
3. Parental SES	1. Father's high-SES (Notary) 0. Father's low-SES (Painter)	Indirectly: (1) Father's contact email (corporate) in the student's file; (2) Father's occupation embedded in the student's essay.
4. Cultural capital	1. High (highbrow culture) 0. Low (popular culture)	Indirectly: Embedded in student's essay.
5. Language ability: essay's objective quality	1. High (good essay) 0. Low (bad essay)	Indirectly: Student's essay
6. Academic performance: subjects failed in the last 6th grade term assessment	1. None 0. Three core subjects	Directly: Student's file academic record
7. Socio-emotional skills	1. Good behavior and high effort 0. Bad behavior and low effort	Directly: Student's file academic record

Figure 1. *Student's file example: Experimentally manipulated factors and levels (in blue) and fixed information (in black) in the vignettes (translated from Spanish)*

SCHOOL DATA		
School name: Pre-primary and Primary Education School Galileo Galilei		***School ID:*** 1400553529
STUDENT'S DATA		
Data of birth: 15/06/2011	***Sex:*** Male / Female	***Nationality:*** SPANISH
Name and Surname(s): Daniel / Lucía García González Youssef / Salma Salhi		
FAMILY DATA		
Father's contact email: Mohamed.Salhi@Painters-Express.es / @Notary-Salhi.es David.Garcia@Painters-Express.es / @Notary-Garcia.es	***Address:*** May 20th Street, 16, 2-B, Madrid	
ACADEMIC RECORD		
Academic year: 2022/2023	***Grade / Term*** 6th grade / 3rd term	
Behavior: Does not respect / respect the classroom norms, exerts low / high effort and motivation and does the homework rarely / most of the time.	***Failed Subjects:*** None / Three core subjects	

3.3.2. Factorial manipulations: measurement and signaling instruments

This study implemented strategies to avoid social desirability bias by hiding the true scope of the research while using signaling instruments as realistic as possible: a table resembling a real student's file and a digitally transcribed essay written by real students. First, as shown in Figure 1 as an example of its implementation in *Qualtrics*, we built a fake but realistic student file, including students' personal information and academic records. We replicated real students' files[11] used by in-service teachers in Spanish primary schools to signal student-ascribed factors to assess discrimination (gender, migrant origin, and family SES) and to provide information on objective ability indicators (academic performance and socio-emotional skills). However, to hide the true aim of the experiment and minimize social desirability, we identified five factors that are usually included in these student's files and kept them constant across respondents: (1) the academic year (2022/2023), term (last term evaluation), and grade (sixth and last grade of elementary education); (2) the student's date of birth and age (15/06/2011; 11-12 years olds born) signaling no previous grade retention; (3) fake school name and administrative ID with a neutral school type signal; (4) fake family address with a neutral signal of the type of house and area; (5) and student Spanish nationality to signal that all students from ethnic minority origin are second-generation. Second, to signal students objective language ability and cultural capital, we used an essay varying by its objective externally validated quality (see online supplement A.4. and Table A.5.) where we subtly embedded cultural capital signals and reinforced again the signals on gender, migration background and family SES previously presented in the student file. Below, we detail how each dimension is operationalized and presented.

Gender

Student names vary by gender and migration background. We select the most common and region-neutral (i.e., no names from bilingual Spanish regions) boy/girl names in the birth cohort of babies born in 2011 (aged 12 in 2023: the average age of a 6th grader), according to the *Spanish Statistics Institute* (INE 2023). For Spanish-origin students, boys are named *Daniel* (0) and girls *Lucía* (1), while for Moroccan-origin (see below) pupils, the boys' name is *Youssef* (0), and girls' is *Salma* (1). Gender is signaled in the student file and the essay's screen instructions.

Migrant origin

We signal migrant origin through first and last names. More specifically, student first names (which vary by ethnic origin and gender) and last names as well as father name and surname. We picked the most common Spanish and foreign-origin surnames among new-borns in 2011 for children and fathers (INE 2023): Spanish origin (1): García; González. Foreign origin (0) - Moroccan: Salhi. For the fathers, we chose the

[11] To preserve the coherence and realism of the student's file structure, we did not randomize the order in which the items are shown across respondents. The table is shown twice to each respondent in two different screens.

most common father's name according to the INE among those born in the 1980s and varied them by ethnic origin: Spanish origin (1): David; Foreign origin (0) - Moroccan: Mohamed. Migration origin is signalled in the student file and the essay's vignette instructions (i.e., the student's name).

Since we could only include two levels for this factor for statistical efficiency, we chose Moroccans as the ethnic minority group of reference for three reasons. First, Moroccans represent the largest share as a foreign-origin minority, with 28.9% (19.6%) of primary (lower-secondary) education students having a Moroccan nationality in 2020-2021 (Ministerio de Educación y Formación Profesional 2023). Second, among the largest ethnic minority groups in the Spanish school system (i.e., in descending order by size: Moroccans, Romanians, Chinese and Latin Americans), Moroccans are the most socioeconomically and academically disadvantaged (Gil-Hernández and Gracia 2018), and also experience the most negative stereotypes (Martínez de la Fuente 2021). Third, Moroccan-origin names and surnames are a more powerful signal identifiable as an ethnic minority in the Spanish context than Romanian or Latin American.

Parental SES

In previous research, students' SES was subtly signaled through names and surnames (Wenz and Hoenig 2020). However, since participants might not easily identify SES variation within foreign origin names and surnames (Crabtree et al. 2022), we signal students' SES through the father's occupation both in the student's file and embedded in the essay (see below). We selected parental (only father's occupation to control for family structure [i.e., single mothers]) occupations commonly perceived as having a high or low SES or prestige according to the ISEI and SIOPS scales (Ganzeboom and Treiman 1996): Low-SES (0) - construction painter (ISCO-08=7131; ISEI=31; SIOPS=29); High-SES (1): notary (ISCO-08=2619; ISEI=82; SIOPS=71). We signaled parental SES subtly in the family contact module of the student's file (see Figure 1) through the father's email (Martínez de la Fuente 2021), including name and surname (as explained above) and business/occupation. For the high-SES occupation (notary), we included the father's surname in the email domain to elicit that the father owns a small-to-medium notary firm. We do this so that participants do not underestimate the SES of ethnic minority fathers compared to the native majority (Crabtree et al. 2022). For the low-SES occupation (painter), we did not include the father's surname in the email domain to elicit that each father is an employee within the firm and will signal in the essay that the father paints houses at work. Low-SES (0): *Moroccan: Mohamed.Salhi@Pintores-Express.es; Spanish: David.Garcia@Pintores-Express.es* High-SES (1): *Moroccan: Mohamed.Salhi@Notarios-Salhi.es; Spanish: David.Garcia@Notarios-Garcia.es*

To further reinforce the SES signal, we also elicit the father's occupation within a sentence embedded in the essay that naturally flows with the paragraph's topic and context: *My family and I love spending time in nature, we all have fun, and my father can disconnect [from painting houses at work / from work at the notary office*] (see online supplement Table A.5. for details).

Cultural capital

We signal the children's embodied cultural capital (Sullivan 2002) within the student's essay (see the placement of cultural capital indicators in the essay in the online supplement Table A.5.). By design, cultural capital is orthogonal to essay objective quality and parental SES (see below). We signal cultural capital through references to student and family participation in highbrow or lowbrow leisure activities that convey different social statuses, recognition, or legitimacy in the dominant cultural hierarchy (Jæger, Rasmussen and Holm 2023; Bourdieu 1984). Low cultural capital (0) is signaled through a lowbrow or popular leisure activity referenced in the essay: watching a reality show on television (i.e., *Temptation Island*) (Childress et al. 2021; Lizardo and Skiles 2009; Bennet 2006). High cultural capital (1) is signaled by a highbrow leisure activity highlighted in the essay: visiting an art museum and knowledge of impressionist paintings (i.e., *Monet*) (Jæger, Rasmussen and Holm 2023). To ensure respondents perceive the embodied cultural capital signals as highbrow or popular culture, we successfully pre-tested its internal validity with 243 in-service elementary education teachers (see online supplement section A.5. for details).

Language ability: objective essay quality

We randomly assigned two versions of a short essay (278-295 words) varying in its objective quality (0=bad; 1=good) regarding structure, orthography, vocabulary, and creativity to capture student's language ability (see online supplement Table A.5. for a transcription). We asked real sixth graders to write essays about a landscape of their liking, a neutral topic, and we adapted them to make them also neutral regarding the urban/rural habitat and region, gender, parental SES, and ethnicity. The essays were digitally transcribed after unsuccessfully pre-testing hand-written versions with 260 in-service elementary education teachers in the Madrid region (i.e., legibility and non-response were issues in the bad version of the essay) (Xu and Gong 2017). As an external benchmark of objective quality (Quinn 2020), we applied official Spanish competence guidelines and rubrics for the 6th grade of elementary education. We pre-tested the objective grade assigned to the digital essay using a sample of 243 in-service elementary education teachers in the Andalusian region. As a result, as illustrated in Figure A.2. of the online supplement (section A.4.), showing the kernel distributions, and in Table 3, in-service teachers assigned a 5.5 (SD=1.4) average grade to the bad essay and 8.9 (SD=1.1) to the good essay on a 1-to-10 scale, where 1 is the lowest and 10 is the highest grade, in accordance with real grading practice, with a joint mean at 7.2 (SD=2.1). Moreover, we asked teachers to assess the essays' degree of credibility (i.e., written by a 6th grader) and guess the gender of the writer. About 60% of respondents reported the essay as credible, and about 70% could not say if a boy or a girl wrote it.

As shown in Table 3 below, our experimental sample, consisting of pre-service teachers enrolled in the BA in Primary Education at a Faculty of Education, assigned a

5.9 (SD=1.5) average grade to the bad essay and 8.7 (SD=1.3) to the good essay on a 1-to-10 scale, with a pooled mean of 7.3 (SD=2). Thus, our measure of student language ability shows high external and internal validity.

Table 3. *Essay grades summary statistics*

Essay	n	Mean	SD	Min	Max
In-service Teachers (pre-test)					
Overall Essay	243	7.2	2.1	1.6	10
Bad Essay	123	5.5	1.4	1.6	8.6
Good Essay	120	8.9	1.1	3.3	10
Pre-service Teachers (experiment)					
Overall Essay	1,717	7.3	2	1	10
Bad Essay	846	5.9	1.5	1.1	10
Good Essay	871	8.7	1.3	1	10

As shown in the online supplement section A.4. and Table A.5., to increase the signaling power of our factorial manipulations, we exploit eight versions of the essay orthogonally varying by its objective quality (2), cultural capital (2), and parental SES (2). To reinforce our signals, we embed the student's name and surname—signaling their gender and ethnic origin—within the essay's screen instructions and the father's occupation within a sentence embedded in the essay.

Academic performance: Number of subjects failed

To account for the student's true academic ability, we signal the number of (non-specified) core subjects (i.e., Math, Spanish, Social or Natural Science, and first foreign language) the student has failed/passed in the last term evaluation of the sixth grade: (0) three (non-specified) core non-passed subjects (around the threshold for non-automatic grade promotion according to Spanish educational law); all subjects passed (1). Student academic performance is signaled in the student file.

Socio-emotional skills

To capture student's socio-emotional or non-cognitive skills, one of the strongest predictors of academic performance that might still influence teacher biases in assessments independently of student scholastic competence (Ferman and Fontes 2023; Owens 2022), we include a dummy variable stating if the student exerts high effort, regularly does the homework and behaves well at the classroom (1), or exerts low effort, rarely does the homework, and misbehaves at the classroom (0). These socio-emotional skills are signaled in the student's file.

Outcomes

All outcomes are measured in a metric scale to maximize variation. Table 4 below reports the summary statistics of the outcomes.

Essay grading comprises a 1-10 scale including decimal points (i.e., in the Spanish educational system, grading with 0 is forbidden in primary education), asked with the following question: *What grade from 1 to 10 (including decimal points) would you give to the essay considering its syntactic structure, orthography, vocabulary, and creativity?*

Grade retention recommendations range from 0 to 10, including decimal points, asked with the following question: *Considering the information in the student's file, the grade you assigned him/her in the essay and that he/she has not repeated a grade before, do you think this student should repeat 6th grade? On the scale where you can include decimal points, 0 means that he/she should never repeat 6th grade and 10 means that he/she should definitely repeat 6th grade.* Grade retention in elementary education is discouraged by Spanish educational authorities, but its prevalence in 2020 was 2.3%, considerably above the OECD average of 1.3%.

Educational expectations about enrolment in the upper-secondary academic track are captured with a 0-10 scale including decimal points, asked with the following question: *Considering the information in the student's file and the grade you assigned him/her in the essay, do you think it is likely that this student will reach the upper-secondary academic track? On the scale where you can include decimal points, 0 means that it is not at all likely to happen and 10 means that it is very likely to happen.* The upper-secondary academic track in the Spanish educational system is a two-year academic pathway giving direct access to college—after passing a standardized national entry exam for public universities. To enrol in upper-secondary education, either in the vocational or academic track, students must get a diploma after passing the 4-grade lower-secondary cycle (ESO).

Table 4. *Outcomes' summary statistics, skewness and kurtosis*

Outcome	n	Mean	q50	SD	Min	Max	Skewness	Kurtosis
Essay Grade	1,717	7.32	7.5	1.97	1	10	-0.46	2.41
Grade Retention	1,717	3.00	2	2.95	0	10	0.65	2.24
Academic Track	1,717	7.29	7.7	2.16	0	10	-0.75	3.07

3.3.3. Estimation and models

In the baseline set of models (M1) (only shown in the online supplement, Tables A.6.-A.8.), we run Ordinal Least Squares (OLS), i.e., linear regressions, including a dummy for each of the seven experimental factors, to estimate their *Average Marginal Component Effect* (AMCE) on the three metric outcomes ($y_{i1essay\ grade}$; $y_{i2retention}$; $y_{i3expectations}$). The AMCE can be interpreted as the causal effect of a specific factor level (i.e., treatment) in comparison with another level of this same factor (i.e., baseline or control category) while keeping equal the joint distribution of the remaining factors, averaged across this distribution and the population's sampling distribution (Hainmueller, Hopkins, and Yamamoto 2014:29). Furthermore, standard errors are clustered at the faculty/university level to account for the non-independence of observations within these sampling clusters.

In the second set of models (M2), as formally presented in Equation 1 below, pre-treatment respondent-level controls (ethnic origin, gender, parental SES, grade retention in primary and/or secondary school, institution fixed effects, BA enrolment grade and year of birth) are a covariate vector ($\boldsymbol{Z}_j$) to account for the (unlikely) potential unbalance or confounding of these individual-level characteristics across our experimental conditions and increase the precision of our main effects of interest, as individual-level characteristics might partially confound them even after the vignettes' randomization (Baguley et al. 2022).

$$y_{i123} = \alpha + \beta_{i1}\, gender + \beta_{i2}\, migrant\ background + \beta_{i3}\, parental\ SES + \beta_{i4}\, cultural\ capital + \beta_{i5}\, essay\ quality + \beta_{i6}\, subjects\ failed + \beta_{i7}\, socioemotional\ skills + \boldsymbol{Z}_j + \varepsilon_i \quad \text{(Equation 1)}$$

Finally, according to the pre-registered power analysis and final analytical sample we reached in the study, we cannot generally estimate moderation analyses by interacting different factors with enough statistical precision. Thus, we do not run further moderation models except for two specific cases relevant to the analyses where large and powered heterogeneous effects can be identified by essay quality (i.e., essay grading outcome) and the number of failed subjects (i.e., grade retention recommendation outcome) (see Sections 4 and 4.1.).

4. Findings

Table 5 portrays the main OLS models (M2) output by each outcome and experimental factor, controlling for sampling institution-fixed effects and respondent characteristics. In Table 5 and Figure 2, upper panel, we split the independent variables by *ascribed* and *ability* factors, representing multiple randomized categorical treatments of each factor level with respect to its baseline or control group (i.e., reference categories in

parentheses in Table 5). The coefficients of these factors or variables are AMCEs, the causal estimand of interest, expressing a factor's average individual-level causal effects that aggregate possible heterogeneous effects over respondents. For instance, the AMCE of an objectively *good* student essay (*treatment*) on the respondent's essay's grading (β_{AMCE} = 2.83) versus an objectively *bad* essay (*control*) is calculated by averaging the grades of all student's profiles with a *good* essay, averaging the grades of all student's profiles with a *bad* essay, and subtracting them. The remaining factors operate as randomized pre-treatment covariates. In this way, we can answer how much the essay grade of a student's profile would change if a student's ability, measured with the objective essay's quality, changed from objectively *bad* to *good*, on average, over the remaining factors' distribution.

Table 5 shows that those factors accounting for students' objective ability are the most predictive *vis-à-vis* ascriptive factors across all three outcomes. That is unsurprising, as students' cognitive and non-cognitive skills are well-known as the leading indicators of educational performance for teachers' assessments. Intuitively, if we focus on the first outcome, essay grading, an objectively *good* essay implies 2.8 points (p-value<0.001) higher teacher grading on a 1-10 scale (σ = 2) than an objectively *bad* essay. The remaining factors accounting for student objective ability, number of failed subjects, and socio-emotional skills (classroom behavior and effort) have similar predictive power, with AMCEs at 0.28 (p-value<0.01) and 0.27 (p-value<0.01), respectively. One should note that, in the specific task of grading an essay, a student's number of failed subjects or classroom behavior is arguably outside of the scope of what should be graded. This finding mirrors previous research showing that teachers unfairly assess students according to their classroom behavior, instead of their objective competence (Ferman and Fontes 2023).

As Figure 2's upper panel illustrates, the relative effect size of the most relevant ability factor for essay grading, objective essay equality, is about 17 times larger (2.832/0.17=16.7) than the average effect size for those statistically significant student-ascribed status characteristics ($\beta_{AMCE} \approx 0.17$). Figure 2's bottom panel zooms in on the particular role of the latter. Net of students' objective observed ability, teachers tend to assign higher grades in the essay, on average, to students profiles who are girls (β_{AMCE} = 0.12 [p-value<0.1]), come from a (Moroccan) ethnic minority origin (β_{AMCE} = 0.2 [p-value<0.01]), or signal high cultural capital (β_{AMCE} = 0.2 [p-value<0.001]).

Concerning this latter finding on cultural capital discrimination validating *hypothesis 3*, it is worth mentioning that this pattern will not replicate in the remaining outcomes on teacher expectations of student educational transitions. Instead, direct exposure to a written highbrow cultural capital signal—i.e., an art museum visit and a metaphor to an abstract painter's works—embedded in a real student's task, an essay, elicits higher teacher evaluations. In doing so, in line with *hypothesis 3*, teachers might misconceive students' high cultural capital with academic brilliance, as these factors are orthogonal by design, independently of a student's SES and objective ability.

In opposition, net of usually correlated factors in real life—a student's objective performance, classroom behavior, ethnic origin, and cultural capital—student parental social class is irrelevant in predicting teacher grades, partially rejecting *hypothesis 1* on implicit bias or SCT. At the same time, though, this null finding is in line with a previous factorial experiment following a similar design as ours (Wenz and Hoenig 2020) and with an observational study on elementary education students in the Spanish Andalusian region (Marcenaro-Gutiérrez and Vignoles 2015). However, it contrasts with observed teacher bias by student SES—i.e., over-grading of high-SES pupils—in the Spanish Basque Country region (Gortázar, Martínez de Lafuente and Vega-Bayo 2022) and other countries. Still, as in most previous observational studies, the latter study did not control for relevant students' behavioral characteristics, suggesting a potential overestimation of SES discrimination due to omitted variable bias (Ferman and Fontes 2023) and/or measurement error (van Huizen, Jacobs and Oosterveen 2024).

For the case of student sex, the magnitude and direction of the coefficient point to slight (6% of a 1-unit SD) positive grade discrimination for girls, as expected by *hypothesis 1* and broadly in line with previous findings. However, the estimation has high uncertainty with a p-value < 0.1, suggesting caution. As highlighted above, the power analysis indicated a minimum detectable effect size (p-value < 0.05) for the essay grading outcome at β_{AMCE} = 0.133, with the gender coefficient in Table 5, M2, just below this threshold at β_{AMCE} = 0.121. Still, given this small margin and the overall positive direction of the gender factor across outcomes, it is likely underpowered due to its small effect size (i.e., false negative). As illustrated by Figure A.7. in the online appendix, looking at heterogeneity in the effect of gender on essay grading reveals that girls are particularly over-graded in comparison with boys when the essay is good, with a considerably larger β_{AMCE} = 0.269 (p-value<0.05; n=871) than the above AMCE over essay quality, backing the generalized belief that girls are more competent than boys in language tasks (Homuth, Thielemann and Wenz 2023). Among bad essays, there are null gender differences in grading at β_{AMCE} = 0.008 (p-value>0.9; n=846).

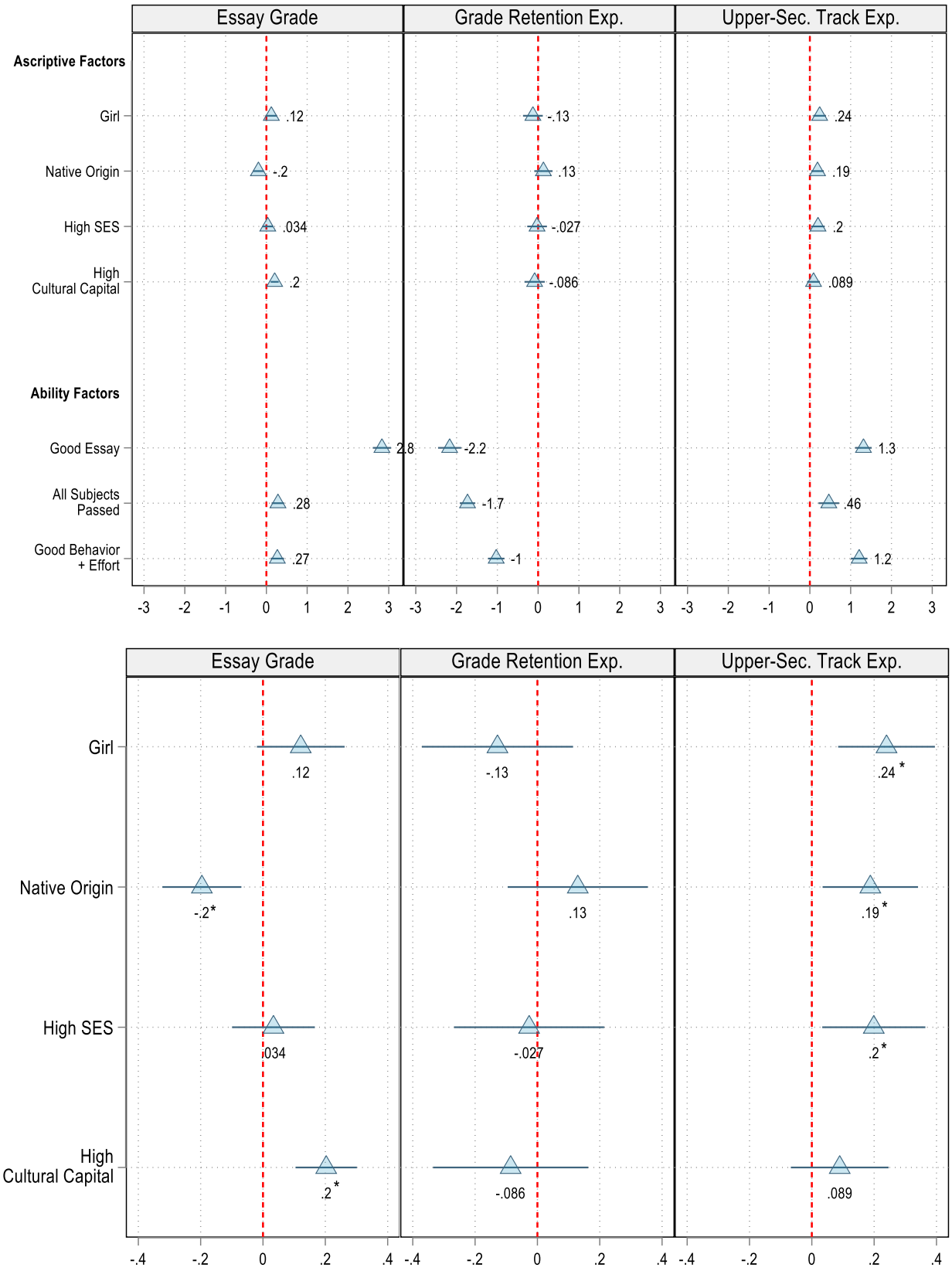
In turn, the finding on positive ethnic discrimination in essay grading, contrary to *hypothesis 1*, is surprising. We did not expect this coefficient's direction favoring ethnic-origin students under any of the above-discussed discrimination theories. Over and above student objective ability and other correlated—in the real world, not in this experimental design—ascribed status characteristics, such as parental class and cultural capital, it seems that teachers tend to compensate for a student's overall disadvantaged ethnic-minority origin by over-grading them (Schuessler and Sønderskov 2023) in comparison to the equally skilled and socio-economically (dis)advantaged ethnic majority (i.e., Spanish origin). We further discuss the interpretation and implications of this unexpected finding in the concluding *section 5* in line with compensatory discrimination theories.

Table 5. *OLS main models (M2): Experimental factors on educational outcomes*

	Outcomes		
Randomized Factors	Essay Grade (1-10)	Grade Retention Recommendation (0-10)	Academic Track Expectations (0-10)
Ascriptive Factors			
Female (Male)	0.121^{+}	-0.128	0.240**
	(0.067)	(0.115)	(0.074)
Native Origin	-0.196**	0.129	0.188*
(Moroccan Origin)	(0.060)	(0.107)	(0.073)
High-SES (Low-SES)	0.0335	-0.0266	0.199*
	(0.063)	(0.115)	(0.079)
High Cultural Capital	0.203***	-0.0859	0.0895
(Low CC)	(0.047)	(0.118)	(0.075)
Ability Factors			
Good Essay (Bad Essay)	2.832***	-2.169***	1.313***
	(0.107)	(0.135)	(0.096)
All Subjects Passed	0.283**	-1.731***	0.465**
(3 Core Subjects Failed)	(0.073)	(0.091)	(0.120)
Good Behavior + Effort	0.268**	-1.027***	1.209***
(Bad Behavior + Effort)	(0.078)	(0.095)	(0.097)
Individual Controls	✓	✓	✓
Institution Fixed Effects	✓	✓	✓
Observations	1,717	1,717	1,717
Adjusted R^2	0.522	0.254	0.186

Notes: Clustered standard errors by institutions in parentheses, individual-level controls: year of birth, gender, country of birth, parental country of birth, parental highest education, BA Degree enrollment grade, grade retention in primary and/or secondary school. Two-tailed t-tests: $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 2. *AMCEs of ascriptive and ability factors on educational outcomes (95% CI)*



Notes: *p-value < 0.05 (Bottom-panel); Controls: institution-FE; respondents' characteristics. The bottom panel displays the same model as the upper, zooming in on ascriptive factors.

We now turn to the second outcome, teacher recommendations for student grade retention in elementary education's 6th (and last) grade. As expected, all ability-related factors are statistically significant and highly predictive of teacher grade retention recommendations. Again, similarly to the essay grading outcome, the objective quality of the essay, as a proxy for the student's true (language-related) ability, is the most predictive factor (β_{AMCE} = -2.16 [p-value<0.001]) of grade retention recommendation. Following in effect size, having passed all subjects (β_{AMCE} = -1.73 [p-value<0.001]), and students' good behavior and effort (β_{AMCE} = -1.03 [p-value<0.001]) in the current term evaluation are considerably more predictive than they were for essay grading. In this case, the larger effect size of the student's socio-emotional skills and, notably, the overall performance (number of subjects passed) aligns with the outcome's nature. Legal thresholds for granting repetition are set at three core failed subjects, and teachers are particularly prone to recommend a student to repeat if he/she does not strive or misbehave as a *punishment policy*.

If we focus on the student's ascribed status characteristics, even though the coefficients' direction and effect sizes generally align with our findings for essay grading, none is statistically significant under the standard 5% threshold. When interpreting these *a priori* null findings on teacher bias or discrimination, thus rejecting *hypothesis 1* for grade retention recommendations, one should consider, however, the outcome's high variation and skewed distribution. As shown above, the power analysis indicated a minimum detectable effect size (p-value < 0.05) for the grade retention recommendation, on a 0-to-10 scale, at β_{AMCE} = 0.199, considerably higher than the one for the essay grading outcome, likely driven by its higher variation at σ = 2.95. Additionally, one should note its highly skewed distribution to the right (mean = 3; median = 2), seemingly indicating that most teachers are highly averse to grade retention.

In addition, we should further consider the specific context of grade retention in the Spanish (primary) educational system regarding prevalence, legal criteria, and social norms. As pointed out above, as in most OECD countries, Spanish educational authorities discourage this practice, especially in primary education, with a low absolute prevalence at 2.3% but high relative to the OECD average of 1.3% and considerable enforcement heterogeneity by regions and schools. For recommending or expecting a student to repeat a grade in primary education, there must be a strong underperformance signal, with legal criteria establishing at least three core failed subjects. Accordingly, among those students' profiles with three core subjects failed, the average (median) values for grade retention recommendation stands at 3.8 (4), considerably higher than for those students' profiles passing all subjects at 2.1 (0.7). Indeed, as shown in the online appendix Figure A.8., the distribution is most balanced in the case of failing three subjects.

Thus, we run a heterogenous model (M2) by the number of failed subjects (none or three core subjects) to mitigate the skewness in the joint distribution and test for a

more realistic setting to improve the external validity of the findings. As illustrated by Figure A.9. in the online supplement, in line with previous findings on essay grading, despite halving the sample size (n = 867), we find strong positive gender (β_{AMCE} for girls = -0.36 [p-value<0.05]) and ethnic-minority (β_{AMCE} for Moroccans = -0.53 [p-value<0.01])—compensatory—discrimination in teacher grade retention recommendations, in line with and contrary to *hypothesis 1*, respectively.

The third outcome is teacher expectations about a student's enrollment in the upper-secondary academic track. As stated in *hypothesis 2*, findings on ascribed status discrimination in this long-term outcome could lean more toward statistical discrimination explanations than implicit bias or SCT compared to the former short-term outcomes, as teachers evaluate students' profiles considering the last grade of elementary education and therefore do not have other necessary information about their potential future performance. Lower-secondary education comprises four grades, if there is no grade retention, before the end of compulsory education (16 years old) and the transition into upper-secondary education, either to 2-year lower vocational tracks or academic tracks leading to college. Thus, as anticipated by *hypothesis 2*, one can expect discrimination effect sizes to be larger than for essay grading or grade retention recommendations and more likely explained by statistical discrimination mechanisms than by implicit bias or SCT.

With this setting in mind, Table 5 and Figure 2 display the output from the main model (M2). As already visible from the smaller adjusted coefficient of determination (R^2) at the bottom of Table 5 compared with essay grading and grade retention, ability factors signaled in elementary education grade 6th—such as student objective language ability (essay quality) and overall performance (number of subjects passed; β_{AMCE} = 0.47 [p-value<0.001]—have smaller effect sizes for this third outcome. Still, student classroom behavior and effort (β_{AMCE} = 1.21 [p-value<0.001]) seem to gain weight compared to the previous two outcomes as a powerful indicator of future success, with a similar effect size as language ability (β_{AMCE} = 1.31 [p-value<0.001]). Overall, these patterns hint that students' current performance is not fully informative for teachers to predict future attainment accurately, leaving room for ascribed status group-level stereotypes, assumptions, and performance to play a non-negligible role.

Focusing now on ascribed status factors reveals that, in line with *hypothesis 2*, statistical discrimination theories and previous experimental and observational findings (Geven et al., 2021; Timmermans, Kuyper and Werf, 2015), teachers express higher long-term expectations for those groups with, on average, historically higher educational performance in lower-secondary education and chances of attending the upper-secondary academic track (Gil-Hernández and Gracia 2018), such as girls (β_{AMCE} = 0.24 [p-value<0.01]), native origin (Spanish origin; β_{AMCE} = 0.19 [p-value<0.05]) and high-SES (β_{AMCE} = 0.2 [p-value<0.05]) students, displaying similar effect sizes at about 10% of a SD-unit (σ = 2.16).

Table 6. *Findings summary and benchmarking with observational research*

	(1) Observational Teacher Bias[b] (SD)[a]	(2) Experimental Teacher Bias[c] (SD)	(3) Experimental / Observational (2/1)*100	(4) Total Observed Gap[d] (SD)	(5) Experimental / Total Gap (2/4)*100	(6) Hypotheses Validation (2)
			Grading			
Girl	0.27**	0.06+ [0.14*]	23 %	0.12***	53 %	✓ (H1)
High-SES	0.22**	0.02	8 %	0.54***	3 %	x (H1)
Native Origin	0.14**	-0.10**	-69 %	0.79***	-13 %	**x** (H1)
High Cultural Capital		0.10***				✓ (H3)
			Grade Retention			
Girl		-0.04 [-0.36*]		-0.08***	55 %	✓ (H1)
High-SES		-0.01		-0.24***	4 %	x (H1)
Native Origin		0.04 [0.53**]		-0.32***	-14 %	**x** (H1)
High Cultural Capital		-0.03				x (H3)
			Educational Expectations			
Girl		0.11**		0.14***	78 %	✓ (H1; H2)
High-SES		0.09*		0.46***	20 %	✓ (H1; H2)
Native Origin		0.09*		0.01	664 %	✓ (H1; H2)
High Cultural Capital		0.04				x (H3)

Notes: [a] SD = Standard Deviation; Blank squares with no available or comparable data in Spain. In column (6), ✓ indicates those statistically significant (p-value < 0.05 with a two-tailed t-test) estimates that partially confirm the article's research hypotheses; X marks those non-statistically significant, null effects or statistically significant coefficients that identify an opposite-sign pattern than expected (additionally in bold) that partially reject the corresponding research hypothesis. H1=Status characteristics and implicit bias theories; H2=Statistical discrimination theory; H3=Cultural reproduction theory. Between brackets are estimates from heterogeneity analyses by students' objective performance (essay quality or number of subjects failed).

[b] Estimates by Gortázar et al. (2022) on the z-standardized difference between teacher's assigned grades and (low-stakes) blind test scores in Spanish with data (n=15,802) from the Basque Country region (Spain) among 4th graders in the courses 2015/16 and 2016/17; High-SES = family Socioeconomic and Cultural Index (difference between 3rd and 1st tercile); native = students with parents born in Spain vs. all 2nd generation migrant-origin students (at least one foreign-born parent).

[c] Estimates from Table 5 (n=1,717) on fictitious students' profiles of 6th graders; OLS models experimentally controlling for student's ability on (pre-service) teachers' grades of a short essay, grade retention recommendations, and educational expectations for enrolment in the academic upper track in secondary education.

[d] Own elaboration with data (n=22,500) from a national evaluation among 4th graders (Ministerio de Educación, 2009); High-SES = skilled workers vs. professionals (fathers); native = students with both parents born in Spain vs. 2nd generation Moroccan-origin students with both parents born in Morocco. OLS and LPM on Spanish standardized blind test scores, grade retention in 2nd or 4th grade, and parental expectations (`*What educational level are you hoping for that your child is studying?'*) for their children's educational attainment (1 = university or academic upper-secondary track; 0 = compulsory education or vocational training) with controls for gender, father's occupation, migrant-origin, and month of birth, with clustered standard errors by schools.

Two-tailed t-tests: + p < 0.10, * p < 0.05, ** p < 0.01, *** p < 0.001

In this case, in comparison with a short-term outcome such as essay grading where teachers are supposed to count with all necessary information for a fair assessment, the relative effect size of the most relevant ability factors for teacher educational expectations—objective essay equality (1.313/0.21=6.3) and student behavior (1.209/0.21=5.8)—is considerably smaller at about 6 times larger than the average effect size of students' statistically significant ascribed characteristics ($\beta_{AMCE} \approx 0.21$). These findings suggest that student-ascribed status might gain weight in teacher evaluations vis-à-vis ability-related factors the less reliable and the less complete information on the latter is.

Furthermore, as predicted by *hypothesis 2*, the effect sizes of gender ($\beta_{AMCE\text{-}Expectations} - \beta_{AMCE\text{-}Grading} = 0.119$ [SE = 0.175]; p-value > 0.1), parental class ($\beta_{AMCE\text{-}Expectations} - \beta_{AMCE\text{-}Grading} = 0.165$ [SE = 0.157]; p-value > 0.1), and ethnic origin ($\beta_{AMCE\text{-}Expectations} - \beta_{AMCE\text{-}Grading} = 0.384$ [SE = 0.157]; p-value < 0.001) are substantially larger, from 2 to 6 times—even changing sign for ethnic origin, for long-term educational expectations than for short-term or simultaneous outcomes to the student's information provided, like essay grading. Nevertheless, as shown above in parentheses, formally examining the differences in the regression coefficients from separate (non-nested) models (Clogg et al. 1995) with a two-tailed *z-test* yields statistically significant results only for ethnic background under the standard 5% threshold, suggesting cautious interpretation.

4.1. Robustness checks and additional analyses

We run several pre-registered robustness checks and additional analyses (see online supplement sections A.9.-A.10.) to assess the credibility of our findings. The article's main findings and conclusions are generally consistent over the following robustness checks. Firstly, in Table A.7., we run additional models controlling for the manipulation checks. We replicate the analyses in a subsample of those respondents who correctly recalled[12] all treatment levels (56.9%; n=977). Second, as a deviation from the pre-analysis plan, we replicate the main models adjusted for calibration weights using raking estimators to adjust for the population shares of the main individual-level socio-demographic variables in Table A.8. Third, given that our primary outcomes are significantly non-normally distributed according to a joint normality test (p-value<0.000) based on skewness and kurtosis (see Table 4 above), we estimate alternative model specifications by dichotomizing the outcomes below/above the median of the scale combined with the implementation of linear probability models (LPM) in Table A.9. Fourth, for the outcome on grade retention, in line with Spanish educational law, one should note that teachers can only recommend students' repetition for those students failing at least three core subjects. Thus, in the online supplement Figure A.9., we display a heterogenous model (M2) by the number of failed subjects (none or three core subjects) to mitigate the skewness in the joint distribution

[12] It could also be the case that not recalling a treatment might proxy for non-discriminatory behavior as respondents might not consider a given student's ascribed factor as a relevant piece of information for their assessments.

and test for a more realistic setting.[13] Fifth, in a third set of models (M3), since we found statistically significant main effects of parental SES, gender, ethnic origins, and cultural capital, we test whether the teachers' perception of parental support available for student's education (0-10 scale)[14] is a mechanism by (1) running OLS models of these factors on parental support (Table A.10.), and (2) using the Karlson-Holm-Breen (KHB) decomposition method to test whether parental support mediates the effect of these experimental factors on the corresponding outcomes (Table A.11.). Parental support only mediates or confound the effect of cultural capital on essay grading at 36% (p-value<0.1).

5. Discussion and conclusion

Fair evaluations are a necessary condition to pursue equal educational opportunity. Teachers are the main evaluators of academic performance and merit in the educational system. Nevertheless, their direct role in reproducing educational inequalities remains poorly understood as previous observational work and the few experimental studies available have yielded inconclusive and often inconsistent. Thus, this article tested if (pre-service) teachers show discrimination in their assessments and expectations as a function of student-ascribed status characteristics.

We framed our research hypotheses from multidisciplinary theories of status characteristics beliefs, implicit bias, statistical discrimination, and cultural reproduction. We analyzed different outcomes over the students' educational careers—essay grading, grade retention recommendations, and expectations about academic track attendance—conveying diverse uncertainty for teacher evaluations to shed light on these theories' predictive power. We conducted a pre-registered full factorial survey experiment with realistic and externally validated instruments (i.e., student's file and essay)—with 128 student profiles—and a representative sampling of Spanish pre-service teachers before exposure to students or the school context. For the first time, this research design allows us to causally disentangle the net effect of different student-ascribed status characteristics—gender, class background, ethnic origin, and cultural capital—on teachers' (biased) assessments.

Table 6 above summarizes the article's main findings and hypotheses validation. Overall, we found teacher biases in (essay) grading favoring girls (supporting *hypothesis 1* on *status characteristics and implicit bias theories*), ethnic minority origin (partially rejecting *hypothesis 1*), and students signaling high cultural capital (partially supporting *hypothesis 3* on *cultural reproduction theory*). Regarding teachers' recommendations about grade retention, findings mirror the direction of the former biases for grading by gender and ethnic origin, except for cultural capital, among low-

[13] Even when the reported interaction effects for gender and ethnic-origin are statistically significant (p-value < 0.01) due to large effect sizes, overall, the power analysis did not yield enough power to run interactions with enough precision at lower effect sizes.

[14] Parental support was asked with the following question: *Considering the information in the student's file, how much interest and support do you think the family shows in the student's education? On the scale, 0 means no interest or support and 10 means a lot of interest and support. You may include decimals.*

performing students falling within the legal threshold for repeating a grade—i.e., failing three core subjects. Finally, regarding teachers' educational expectations of enrolment in the academic track leading to college, we found hints of *statistical discrimination* in favor of girls, native origin, and high-SES background students, validating *hypothesis 2*.

Regarding essay grading as a short-term outcome where we allegedly provided teachers with the minimum necessary information for fair assessments and experimentally controlled the objective essay quality, we interpret findings in line with our pre-registered hypothesis framed in theories of *implicit bias* or *status characteristics beliefs* (gender) and *cultural reproduction* (cultural capital signals). The former finding aligns with the generalized belief that girls are more competent at school, as they overperform boys both in achievement and attainment, especially in language competencies. The latter finding on cultural capital, which does not hold for long-term expectations, suggests that exposure to a highbrow culture signal in academic tasks (i.e., an essay) boosts teacher perceptions of academic brilliance and ratings independently of a student's true ability. Cultural capital is orthogonal to student SES by design in our experiment, but given that they positively associate in reality, the former might be a causal mechanism driving SES-based inequality in assessments.

Contrary to our expectations and *hypothesis 1*, there is a null result on parental class in short-term assessments. This null finding aligns with a similar previous factorial experiment in Germany (Wenz and Hoenig 2020) and observational research in Spain (Marcenaro-Gutiérrez and Vignoles 2015). At the same time, it suggests (1) potential overestimation in those studies detecting bias against low-SES students (Gortázar, Martínez de Lafuente and Vega-Bayo 2022) for not fully controlling for socio-emotional skills (Ferman and Fontes 2023) and/or measurement error in test scores (van Huizen, Jacobs and Oosterveen 2024); (2) underestimation in our essay grading task due to low ecological validity since, in the school context, teacher biases might accumulate over several assessments during the whole academic year; or (3) the cultural capital mechanism fully accounting for observed assessment bias by SES.

In turn, we found unexpected evidence of *over-grading* and *under-expectations* of grade retention for ethnic-minority students, which aligns with explicit *compensating discrimination* (Schuessler and Sønderskov 2023). In the egalitarian context of Denmark, Schuessler and Sønderskov (2023) found that teachers tend to overgrade ethnic-minority-origin students if they underperform relative to their national-origin classmates due to (assumed) teachers' equalizing preferences. In our investigation, absolute grading practices should prevail (Hjorth-Trolle, Rosenqvist and Hed 2022) since each respondent only evaluated one student profile. Still, even though student academic performance and SES are orthogonal to ethnic origin by design, teachers might generally perceive that Moroccans underperform compared to the Spanish-origin majority, as the former group is one of the worst-performing minorities in the Spanish educational system (Gil-Hernández and Gracia 2018). Furthermore, about 79% of

second-generation Moroccan-origin students do not regularly speak Spanish at home (Gil-Hernández and Gracia 2018:594). Thus, teachers might generally perceive that they are a disadvantaged minority that might experience language difficulties and thus explicitly compensate for that disadvantage by over-grading. Relatedly, Alesina et al. (2018) found that teachers' negative stereotypes towards migrant-origin students, captured with the IAT test, do not impact their average Italian grades, while they do affect math. The authors argue that this pattern possibly indicates that literature teachers internalize the need to help immigrants less acquainted with the Italian language, regardless of their biases (Alesina et al. 2018:3).

At the same time, the observed biases in educational expectations of upper-secondary pathways—a long-term outcome lacking information on students' future performance—favoring girls, native origin and high-SES students lean more into *statistical discrimination theories*, validating our related *hypothesis 2*, and in line with previous experimental findings (Geven et al. 2021 for SES; Wenz and Hoenig 2020 for SES and migrant origin). Generally, we found effect sizes at least double those identified for essay grading and grade retention recommendations, concurrent outcomes to the student's information provided. These findings also align with observed teachers' stereotypes of students' group-level competencies (i.e., by gender, social origin and migration background) (Homuth, Thielemann and Wenz 2023).

The finding on (negative) discrimination by ethnic-minority origin, which dramatically changes its effect size and direction from positive to negative compared to the remaining outcomes, is particularly striking given the general *optimism* of migrant-origin families and students when expressing their educational expectations (Gil-Hernández and Gracia 2018) and their actual more ambitious enrolment choices (Ferrera 2023), compared to equally-performing peers from national majority origin. Thus, teachers' statistical discrimination practices might lead to self-fulfilling prophecies or adverse P*ygmalion effects* if teachers expect less academic success from those historically disadvantaged or discriminated groups, such as migrant-origin and low-SES students, risking to rationalize stereotypes and legitimize ascribed status inequalities in the name of efficiency (Tilcsik 2020).

On average, as shown in Table 6 above, we reported effect sizes (on average, Cohen's D ≈ 0.1 or 10% an SD) that closely resemble some previous observational studies in Denmark (Schuessler and Sønderskov 2023) and Italy (Alesina et al. 2018), but are considerably smaller than the most comparable observational study carried out in Spain (i.e., Basque Country region) to date (Gortázar et al. 2022). Thus, observational studies might overestimate teacher bias when not accounting for measurement error in test scores and/or not controlling for non-cognitive ability measures (van Huizen, Jacobs and Oosterveen 2024). Ideally, to accurately identify teacher biases with observational data, one should exploit residual differences between fully comparable high-stakes blindly-assigned test scores and teacher-assigned grades covering the

same curricula (Ferman and Fontes 2023; Schuessler and Sønderskov 2023; Bygren, 2020).

One might still wonder to what extent our identified experimental average effect size of teacher bias is substantial as a mechanism of educational inequality relative to the more considerable weight of students' objective ability or ascribed inequalities in its development. As illustrated in Table 6 above, to benchmark our experimental estimates, we calculated gaps in a blind standardized test of Spanish competencies, grade retention, and educational expectations in a nationwide evaluation of 4th graders by—simultaneously controlling for—student's gender, migrant origin (second-generation Moroccan or Spanish) and parental occupation (skilled workers or professionals), mimicking as much as possible our experimental design. For instance, our reported experimental estimates of teacher bias account for more than 50% of the observed gender gaps across all three outcomes. We can also benchmark our average discrimination effect sizes at 0.1 SD with mean learning gains over a school year, from 0.15 to 0.21 SD of literacy ability, or large-scale educational interventions, reporting test scores increases between 0.17 and 0.47 SD (Evans and Yuan 2019). These benchmarks indicate that our identified effect sizes on ascribed status discrimination are not trivial and might entail real consequences for educational pathways, especially when accumulating several assessments (dis)advantages over time (DiPrete and Eirich 2006). For instance, students from disadvantaged backgrounds are generally less risk-averse to downward mobility and have less perceived chances of success in education than advantaged peers (Breen and Goldthorpe 1997). Hence, they may be sensitive to distorting biases in the signaling information teachers' evaluations provide (Holm et al. 2019), potentially pushing their educational expectations downwards. That might be especially prominent among low-performing disadvantaged students around a pass or fail grade, with unclear information on potential success.

This study has four limitations that pave the way for improvements in future research. First, we have done our best to design survey instruments as realistic as possible to emulate real-world evaluation settings (i.e., real essay task and student's file), externally validated with pre-tests applied to in-service teachers. Ecological validity is a recurrent concern in factorial survey experiments, but as shown by Krolak-Schwerdt et al. (2017), vignettes of fictitious students yield ecologically valid results of teachers' assessments in real classrooms. Still, a complex trade-off exists between avoiding social desirability and ensuring that respondents internalize the experimental manipulations in factorial designs. Besides, in the actual school context, teachers tend to weigh several assessments over the academic year, grading *on a curve* or relative classroom-level scales, while our vignette experiment induced absolute grading in a single task. However, absolute and relative grading scales might have different implications for students' ascribed status inequalities depending on school segregation or composition (Hjorth-Trolle, Rosenqvist and Hed 2022), while teachers' biases might accumulate over several evaluations to assign the final grade. Field school experiments combining administrative data on fully comparable internal and external grades and

automated cognition tests represent a promising path to overcome these challenges (Alesina et al. 2018).

Second, our sample of pre-service teachers might raise further issues about external validity, as most did not have direct contact with students or the school context yet. Nevertheless, in this study, we identified effect sizes virtually identical to previous observational and experimental studies with in-service teachers. Furthermore, previous research suggests that pre-and in-service teachers exhibit similar bias towards minority students, with no significant differences based on school context or inter-group exposure (Pit-ten Cate and Glock 2019). In line with this finding and in opposition to *contact* and *conflict theories*, a field experiment on ethnic discrimination among Hungarian students (Elwert, Keller and Kotsadam 2023) indicated that randomly manipulating inter-ethnic exposure or ethnic composition within classrooms did not affect discrimination. Accordingly, large-scale observational studies using administrative data in Denmark (Schuessler and Sønderskov 2023) and Italy (Lievore and Triventi 2023) showed that teacher exposure to migrants and teacher's characteristics like gender and migration background do not moderate biases. Likewise, Starck et al. (2020) have shown that American teachers are not different in terms of implicit and explicit racial and pro-White biases in comparison with the general non-teacher population, putting into question the role of schools embracing racial equity and the need for further teacher training to prevent discrimination. To encourage replication, we made this study code and data available so that future studies can test whether our findings generalize to other national contexts or replicate with in-service teachers, testing *inter-group relations* theories.

Third, we applied a rigorous random sampling design to draw a representative sample of the frame population that allowed us to reach a larger analytical sample size (n=1,717) than most previous experimental studies on discrimination in education. Furthermore, we pre-registered a power plan to identify powered effects and bypass most previous underpowered studies from convenience samples. Still, given the small magnitude of the effect sizes identified and the substantial variation of the outcomes, we could not reliably estimate interactions between our analyzed ascribed characteristics to explore intersectionality. Hence, we recommend that future studies collect larger samples, given the benchmark effect sizes and power we reported in this study, to more reliably identify potential false negatives and interaction effects.

Fourth, with our factorial design, we cannot causally identify the relative explanatory ability of the different theories and mechanisms at work, as we did not randomly assign different degrees of student information to teachers or deploy behavioral tests of automatic cognition, like the IAT, to disentangle implicit bias and SCT from statistical discrimination mechanisms directly (Melamed et al. 2019). Nevertheless, by comparing the relative impact of fixed—ascribed and ability—student information by each respondent that varies in the degree of connection, usefulness and uncertainty to evaluate different educational outcomes from elementary to upper-secondary

education, we can indirectly infer which of these complementary theories or mechanisms are more likely to be at play. Future studies had better apply more fine-grained experimental designs to untangle these mechanisms causally and with more statistical power than ours. That is not an easy task since no comparable and validated automatic cognition tests exist to measure biases according to all the ascribed status characteristics analyzed here, and different mechanisms of implicit bias, SCT, cultural capital reproduction and statistical discrimination might operate simultaneously.

Having acknowledged these limitations, we showed for the first time the causal effect of several ascribed status characteristics—gender, class background, ethnic origin, and cultural capital—among equally-competent students on (pre-service) teacher's biased assessments. We uncovered complex dynamics of bias that helped us expand our knowledge of discrimination as a relevant mechanism behind educational inequalities. Consciously or not, teachers perceived some groups of students as more competent, deserving, or likely to succeed despite equal objective performance depending on their ascribed status or cultural capital. That leads to biased assessments in a fictitious experimental setting that might translate into self-fulfilling prophecies and cumulative (dis)advantages over the actual educational system. We also uncovered teachers' *compensating* grading practices favoring migrant-origin students who, in the real world, generally underperform and come from disadvantaged backgrounds. This pattern entails that previously identified teachers' implicit biases against immigrants might not align with their explicit judgment behavior. We are confident that our findings on the roots of teacher bias can contribute to promoting fair evaluations and designing appropriate policy instruments to minimize discrimination during teacher training and school practice.

6. References

Aigner, D., and Cain, G. 1977. Statistical theories of discrimination in labor markets. *Industrial and Labor Relations Review*, 30(2), 175–187.

Alesina, A, M Carlana, E La Ferrara and P Pinotti. 2018. "Revealing stereotypes: Evidence from immigrants in schools", *NBER Working Paper*, 25333.

Almlund, M., A.L. Duckworth, J.J. Heckman and T. Kautz. 2011. 'Personality psychology and economics', in E. Hanushek, S. Machin, and L. Woessmann (eds.), *Handbook of the Economics of Education*, Amsterdam: Elsevier, pp. 1–181.

Arkes, H. R., and Tetlock, P. E. 2004. Attributions of implicit prejudice, or" would Jesse Jackson'fail'the Implicit Association Test?". *Psychological inquiry*, *15*(4), 257-278.

Arrow, K. J. 1973. The theory of discrimination. In O. Ashenfelter and A. Rees (Eds.), *Discrimination in labor markets* (pp. 3–33). Princeton, NJ, Princeton University Press.

Arrow, K. J. 1998. "What Has Economics to Say about Racial Discrimination?" *The Journal of Economic Perspectives*, *12*(2):91–100. http://www.jstor.org/stable/2646963

Aschaffenburg, K., and Maas, I. 1997. Cultural and educational careers: The dynamics of social reproduction. *American Sociological Review*, 573-587.

Auspurg, K., and Hinz, T. 2015. *Factorial survey experiments*. Sage Publications, Thousand Oaks, CA.

Auwarter, Amy E. and Aruguete, Mara S. 2008. "Effects of Student Gender and Socioeconomic Status on Teacher Perceptions." *The Journal of Educational Research* 101(4):242-246, DOI: 10.3200/JOER.101.4.243-246

Baguley, T., Dunham, G., and Steer, O. 2022. Statistical modelling of vignette data in psychology. *British Journal of Psychology*, 113, 1143– 1163. https://doi.org/10.1111/bjop.12577

Bansak, K., Hainmueller, J., Hopkins, D., and Yamamoto, T. 2021. Conjoint Survey Experiments. In J. Druckman and D. Green (Eds.), *Advances in Experimental Political Science* (pp. 19-41). Cambridge: Cambridge University Press. doi:10.1017/9781108777919.004

Batruch, A., Geven, S., Kessenich, E., and van de Werfhorst, H. G. 2023. Are tracking recommendations biased? A review of teachers' role in the creation of inequalities in tracking decisions. *Teaching and Teacher Education*, 123, 103985. https://doi.org/10.1016/j.tate.2022.103985

Bennett, T. 2006. Distinction on the box: Cultural capital and the social space of broadcasting, *Cultural Trends*, 15:2-3, 193-212, DOI: 10.1080/09548960600713080

Berger Joseph, Fisek Hamit, Norman Robert, Zelditch Morris Jr. 1977. *Status Characteristics and Social Interaction: An Expectation-States Approach.* Santa Barbara, CA: Greenwood Publishing Group.

Bernstein, B. 1961. 'Social class and linguistic development: A theory of social learning', in A.H. Halsey, J. Floud and C.A. Anderson (eds.), *Education, Economy and Society*, New York, NY: Free Press, pp. 288– 314.

Borghans, L., A.L. Duckworth, and J.J. Heckman. 2008. 'The economics and psychology of personality traits', *Journal of Human Resources*, 43, 972–1059.

Borjas, G. J., and Goldberg, M. S. 1978. Biased screening and discrimination in the labor market. *The American Economic Review*, 68(5), 918-922.

Botelho, F., Madeira, R. A., and Rangel, M. A. 2015. Racial discrimination in grading: Evidence from Brazil. American Economic Journal. *Applied Economics*, 7(4), 37-52.

Bourdieu, Pierre. 1977. "Cultural Reproduction and Social Reproduction." Pp. 487–511 in *Power and Ideology in Education*, edited by Jerome Karabel and Albert H. Halsey. New York: Oxford University Press.

Bourdieu, P. 1984. *Distinction: A Social Critique of the Judgement of Taste*. Cambridge, Mass: Harvard University Press.

Bourdieu, Pierre, and Jean-Claude Passeron. 1990. *Reproduction in Education, Society and Culture*. London: Sage.

Breen, R. and Goldthorpe, J. H. 1997. Explaining Educational Differentials: towards a Formal Rational Action Theory. *Rationality and Society*, 9(3): 275- 305.

Breen, R., and Jonsson, J. O. 2005. Inequality of opportunity in comparative perspective: Recent research on educational attainment and social mobility. *Annu. Rev. Sociol.*, 31, 223-243.

Breinholt, A., and Jæger, M. M. 2019. How does cultural capital affect educational performance: Signals or skills? *The British Journal of Sociology*, *71*(1), 28-46. https://doi.org/10.1111/1468-4446.12711

Bygren, M. 2020. Biased grades? Changes in grading after a blinding of examinations reform, *Assessment and Evaluation in Higher Education*, 45:2, 292-303, DOI: 10.1080/02602938.2019.1638885

Calarco, J.M. 2014. 'Coached for the classroom: Parents' cultural transmission and children's reproduction of educational inequalities', *American Sociological Review*, 79, 1015–1037.

Carlana, Michela. 2019. "Implicit stereotypes: Evidence from teachers' gender bias. *The Quarterly Journal of Economics* 134(3):1163–1224. Publisher: Oxford University Press.

Carlana, Michela, Eliana La Ferrara, and Paolo Pinotti. 2022. "Implicit Stereotypes in Teachers' Track Recommendations." *AEA Papers and Proceedings*, 112: 409-14**.**

Cea D'Ancona, M. Á. 2016. Immigration as a threat: Explaining the changing pattern of xenophobia in Spain. *Journal of International Migration and Integration*, *17*, 569-591.

Childress, C., Baumann, S., Rawlings, C., and Nault, J.-F. 2021. Genres, Objects, and the Contemporary Expression of Higher-Status Tastes. *Sociological Science*, *8*, 230–264. https://doi.org/10.15195/v8.a12

Chmielewski, A. K. 2019. "The Global Increase in the Socioeconomic Achievement Gap, 1964 to 2015." *American Sociological Review*. https://doi.org/10.1177/0003122419847165

Clogg, C. C., et al. 1995. Statistical Methods for Comparing Regression Coefficients Between Models. *American Journal of Sociology*, *100*(5), 1261-1293. https://doi.org/2782277

Correll, S.J. and Benard, S. 2006. "Biased estimators? Comparing status and statistical theories of gender discrimination", Thye, S.R. and Lawler, E.J. (Ed.) Advances in Group Processes (Advances in Group Processes, Vol. 23), Emerald Group Publishing Limited, Leeds, pp. 89-116. https://doi.org/10.1016/S0882-6145(06)23004-2

Correll, J., and Ridgeway, L. 2006. Expectation states theory. In J. Delamater (Ed.), *Handbook of social psychology* (pp. 29–51). New York, NY: Kluwer Academic and Plenum Publishers.

Crabtree et al. 2022. Racially Distinctive Names Signal Both Race/Ethnicity and Social Class. *Sociological Science*. 10.15195/v9.a18

DiMaggio, P. 1982. Cultural capital and school success: The impact of status culture participation on the grades of US high school students. *American Sociological Review*, 189-201.

DiMaggio, P. 1997. Culture and cognition. *Annual Review of Sociology*, 23(1), 263-287.

DiPrete, T. A., and Buchmann, C. 2013. *The rise of women: The growing gender gap in education and what it means for American schools*. Russell Sage Foundation.

DiPrete, T. A., and Eirich, G. M. 2006. Cumulative Advantage as a Mechanism for Inequality: A Review of Theoretical and Empirical Developments. *Annual Review of Sociology*, 32: 271-297. https://doi.org/10.1146/annurev.soc.32.061604.123127

Downey, D. B., and Condron, D. J. 2016) "Fifty Years since the Coleman Report. *Sociology of Education*." 89(3). https://doi.org/10.1177/0038040716651676

Dziak, J. J., Collins, L. M., and Wagner, A. T. 2013. *FactorialPowerPlan SAS macro suite users' guide* (Version 1.0). University Park: The Methodology Center, Penn State. Retrieved from http://methodology.psu.edu

Elwert, Felix, Tamás Keller and Andreas Kotsadam. 2023. "Rearranging the Desk Chairs: A Large Randomized Field Experiment on the Effects of Close Contact on Interethnic Relations." *American Journal of Sociology*. Forthcoming.

Evans, D., and Yuan, F. 2019. "Equivalent Years of Schooling: A Metric to Communicate Learning Gains in Concrete Terms." World Bank Working Paper No. WPS8752. *The World Bank*. Retrieved from http://documents.worldbank.org/curated/en/123371–550594320297.

Farkas, G. 2003. Cognitive skills and noncognitive traits and behaviors in stratification processes. *Annual Review of Sociology*, 29, 541–562. doi:10.1146/annurev.soc.29.010202.100023

Fazio, R. H., Samayoa, J. A. G., Boggs, S. T., and Ladanyi, J. 2023. "Implicit Bias: What is it?" in *The Cambridge Handbook of Implicit Bias and Racism* edited by Krosnick, JA, Tobias, H., and Scott, AL. Cambridge, England: Cambridge University Press.

Ferman, B. and Fontes, L.F. 2023. Assessing Knowledge or Classroom Behavior? Evidence of Teachers' Grading Bias. *Journal of Public Economics*. https://doi.org/10.1016/j.jpubeco.2022.104773

Ferrara, A. 2023. Aiming too high or scoring too low? Heterogeneous immigrant–native gaps in upper secondary enrollment and outcomes beyond the transition in France. *European Sociological Review*, *39*(3), 366-383. https://doi.org/10.1093/esr/jcac050

Foley, W. (2023) "Status beliefs negatively affect expected university attainment of lower class students", *Education Inquiry*, DOI: 10.1080/20004508.2023.2296143

Foschi, M. (2000). Double standards for competence: Theory and research. *Annual Review of Sociology*, *26*, 21-42. doi:10.1146/annurev.soc.26.1.21

Freitag, Markus and Julian Schuessler. 2020. "cjpowR – A Priori Power Analyses for Conjoint Experiments," R Package.

Ganzeboom HBG, Treiman DJ. 1996. Internationally comparable measures of occupational status for the 1988 international Standard Classification of Occupations. *Social Science Research*, 25,201-239.

Geven, S., Wiborg, O., Fish, R., and van de Werfhorst, H.G. 2021. How teachers form future expectations for students: a comparative factorial survey experiment in New York, Amsterdam and Oslo. *Social Science Research*. Online first.

Gil-Hernández, C. J., and Gracia, P. 2018. Adolescents' educational aspirations and ethnic background: The case of students of African and Latin American migrant origins in Spain. *Demographic Research*, *38*, 577–618. http://www.jstor.org/stable/26457057

Gilgen, S. and Stocker, M. 2022. Discrimination at the Crossroads? Evidence from a Factorial Survey Experiment on Teacher's Tracking Decisions. *Swiss Journal of Sociology*, 48(1) 77-105.

Glock, S., and Klapproth, F. 2017. Bad boys, good girls? Implicit and explicit attitudes toward ethnic minority students among elementary and secondary school teachers. *Studies in Educational Evaluation*, *53*, 77-86. https://doi.org/10.1016/j.stueduc.2017.04.002

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. *Soc Psychol Educ* **17**, 589–607 (2014). https://doi.org/10.1007/s11218-014-9266-6

Glock, S., Krolak-Schwerdt, S., and Pit-ten Cate, I. M. 2015. "Are school placement recommendations accurate? The effect of students' ethnicity on teachers' judgments

and recognition memory." *European Journal of Psychology of Education* 30(2):169-188. https://doi.org/10.1007/s10212-014-0237-2

Goldthorpe, J. H. 2007. "Cultural Capital": Some Critical observations. *Sociologica*, 1(2).

Gortázar, L., Martínez de Lafuente, D., and Vega-Bayo, A. 2022. Comparing teacher and external assessments: Are boys, immigrants, and poorer students undergraded? *Teaching and Teacher Education*, 115, 103725. https://doi.org/10.1016/j.tate.2022.103725

Greenwald, A. G., and Banaji, M. R. 1995. Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, 102(1), 4.

Greenwald, A. G., and Krieger, L. H. 2006. *Implicit bias: Scientific foundation*s. California law review, 94(4), 945-967.

Hainmueller, J., Hopkins, D., and Yamamoto, T. 2014. Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments. *Political Analysis*, 22(1), 1-30. doi:10.1093/pan/mpt024

Hanna, R. N., and Linden, L. L. 2012. Discrimination in grading. *American Economic Journal: Economic Policy*, 4(4), 146-168.

Heath, A., and Brinbaum, Y. 2007. Guest editorial: Explaining ethnic inequalities in educational attainment. *Ethnicities*, 7(3), 291-304.

Hjorth-Trolle, A., Erik Rosenqvist, and Anders Hed. 2022. Grading Practices and the Social Gradient in GPA: Quasi-Experimental Evidence from Sweden, *European Sociological Review*, 38(3): 455–471, https://doi.org/10.1093/esr/jcab053

Holm, A., Anders Hjorth-Trolle, and Mads Meier Jæger. 2019. "Signals, Educational Decision-Making, and Inequality." *European Sociological Review*, 35(4):447–460

Homuth, C., Thielemann, J., and Wenz, S. E. 2023. Measuring Elementary School Teachers' Stereotypes in the NEPS SC2 (NEPS Survey Paper No. 108). Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://doi.org/10.5157/NEPS:SP108:1.0

INE. 2020. *Encuesta de Inserción Laboral de los Titulados Universitarios EILU-2019*. Madrid: Instituto Nacional de Estadística.

INE. 2023. *Estadística del Padrón continuo 2011*. Madrid: Instituto Nacional de Estadística.

INE. 2023. *Estadística de nacimientos 2011*. Madrid: Instituto Nacional de Estadística.

Jackson, Michelle. 2013. *Determined to succeed? Performance versus choice in educational attainment*. Stanford: Stanford University Press.

Jæger, M. M. 2011. Does cultural capital really affect academic achievement? New evidence from combined sibling and panel data. *Sociology of Education*, 84(4), 281–298. doi:10.1177/003804071141701

Jæger, M. M. 2022. "Cultural capital and educational inequality: an assessment of the state of the art". Pp. 121-134 in *Handbook of Sociological Science*, edited by Gërxhani, K., de Graaf, N. D., and Raub, W. Edward Elgar Publishing.

Jæger, M. M., and Breen, R. 2016. A Dynamic Model of Cultural Reproduction. American *Journal of Sociology*, 121(4), 1079–1115.

Jæger, M. M., and Møllegaard, S. 2017. Cultural capital, teacher bias, and educational success: New evidence from monozygotic twins. *Social Science Research*, *65*, 130-144. https://doi.org/10.1016/j.ssresearch.2017.04.003

Jæger, M. M., Rasmussen, R. H., and Holm, A. 2023. What cultural hierarchy? Cultural tastes, status and inequality. *The British Journal of Sociology*, 1– 17. https://doi.org/10.1111/1468-4446.13012

Jennings, J. L., and DiPrete, T. A. 2010. "Teacher Effects on Social and Behavioral Skills in Early Elementary School." *Sociology of Education* 83(2). https://doi.org/10.1177/0038040710368011

Kao, G., and Thompson, J. S. 2003. Racial and ethnic stratification in educational achievement and attainment. *Annual Review of Sociology*, 29(1), 417-442.

Kaufman, J., and Gabler, J. 2004. Cultural capital and the extracurricular activities of girls and boys in the college attainment process. *Poetics*, 32(2), 145-168.

Kingston, P. W. 2001. The unfulfilled promise of cultural capital theory. *Sociology of Education*, 88-99.

Kisfalusi, D., Janky, B., and Takács, K. 2021. "Grading in Hungarian Primary Schools: Mechanisms of Ethnic Discrimination against Roma Students." *European Sociological Review*, *37*(6), 899-917. https://doi.org/10.1093/esr/jcab023

Kisfalusi, D., Janky, B., and Takács, K. 2018. Double Standards or Social Identity? The Role of Gender and Ethnicity in Ability Perceptions in the Classroom. *The Journal of Early Adolescence*, *39*(5). https://doi.org/10.1177/0272431618791278

Kisida, B., Greene, J. P., and Bowen, D. H. 2014. Creating cultural consumers: The dynamics of cultural capital acquisition. *Sociology of Education*, 87(4), 281-295.

Krkovic, K., et al. 2014. Teacher evaluation of student ability: what roles do teacher gender, student gender, and their interaction play?, *Educational Research*, 56:2, 244-257, DOI: 10.1080/00131881.2014.898909

Krolak-Schwerdt, S., Thomas Hörstermann, Sabine Glock and Ines Böhmer. 2017. Teachers' Assessments of Students' Achievements: The Ecological Validity of Studies Using Case Vignettes, *The Journal of Experimental Education*, DOI: 10.1080/00220973.2017.1370686

Lamont, Michele, and Annette Lareau. 1988. "Cultural Capital: Allusions, Gaps and Glissandos in Recent Theoretical Developments." *Sociological Theory* 6:153–68.

Lamont, M. and Small, M. L. 2008. 'How Culture Matters, Enriching Our Understandings of Poverty'. In Harris, D. and Lin, A. (eds) *The Colors of Poverty, Why Racial and Ethnic Disparities Persist*, New York, NY, Russell Sage Foundation, pp. 76 –102.

Lamont, M., Beljean, S., and Clair, M. 2014. What is missing? Cultural processes and causal pathways to inequality. *Socio-Economic Review*, 12(3), 573-608.

Lareau, A. 2011. *Unequal Childhoods. Class, Race, and Family Life. With an Update a Decade Later*, Oakland, CA: University of California Press.

Lehmann-Grube, S.K., Tobisch, A. and Dresel, M. 2023. Changing preservice teacher students' stereotypes and attitudes and reducing judgment biases concerning students of different family backgrounds: Effects of a short intervention. *Soc Psychol Educ* . https://doi.org/10.1007/s11218-023-09862-3

Lievore, I. and Moris Triventi. 2023. Do teacher and classroom characteristics affect the way in which girls and boys are graded?, *British Journal of Sociology of Education*, 44:1, 97-122, DOI: 10.1080/01425692.2022.2122942

Lizardo, O., and Skiles, S. 2009. Highbrow omnivorousness on the small screen?: Cultural industry systems and patterns of cultural choice in Europe. *Poetics*, *37*(1), 1-23. https://doi.org/10.1016/j.poetic.2008.10.001

Lorenz, G., Kogan, I., Gentrup, S., and Kristen, C. 2023. Non-native Accents among School Beginners and Teacher Expectations for Future Student Achievements. *Sociology of Education*. https://doi.org/10.1177/00380407231202978

Marcenaro-Gutiérrez, O., Prieto-Latorre, C., and Sánchez Rodriguez M.I. 2023. Gender differences between teachers' assessments and test-based assessments. Evidence from Spain, *Assessment in Education: Principles, Policy and Practice*, 30:3-4, 320-345, DOI: 10.1080/0969594X.2023.2251715

Marcenaro-Gutiérrez, O. and Anna Vignoles. 2015. A comparison of teacher and test-based assessment for Spanish primary and secondary students, *Educational Research*, 57:1, 1-21, DOI: 10.1080/00131881.2014.983720

Martínez de Lafuente, D. 2021. Cultural assimilation and ethnic discrimination: an audit study with schools. *Labour Economics*, 72, 102058.

Meissel, K., Meyer, F., Yao, E. S., and Rubie-Davies, C. M. 2017. "Subjectivity of teacher judgments: Exploring student characteristics that influence teacher judgments of student ability." *Teaching and Teacher Education 65*: 48-60. https://doi.org/10.1016/j.tate.2017.02.021

Melamed, D., Munn, C. W., Barry, L., Montgomery, B., and Okuwobi, O. F. 2019. Status characteristics, implicit bias, and the production of racial inequality. *American Sociological Review*, 84(6), 1013-1036.

Merton Robert K. 1968. "The Self-Fulfilling Prophecy." Pp. 475–91 in *Social Theory and Social Structure*, edited by Merton R. K. New York: Simon and Schuster.

Mickelson, R. A. 1989. Why does Jane read and write so well? The anomaly of women's achievement. *Sociology of Education*, 47-63.

Miles, A., Charron-Chénier, R., and Schleifer, C. 2019. Measuring Automatic Cognition: Advancing Dual-Process Research in Sociology. *American Sociological Review*. https://doi.org/10.1177/0003122419832497

Ministerio de Universidades, Gobierno de España. 2023. *Datos y cifras del sistema universitario español, publicación 2022-2023*. Madrid: Secretaría General Técnica del Ministerio de Universidades.

Mitchell, G., and Tetlock, P. E. 2017. "Popularity as a poor proxy for utility: The case of implicit prejudice." *Psychological science under scrutiny: Recent challenges and proposed solutions* 164-195.

OECD. 2020. "The PISA target population, the PISA samples and the definition of schools", in *PISA 2018 Results (Volume II): Where All Students Can Succeed*, OECD Publishing, Paris.

Owens, J. 2022. Double Jeopardy: Teacher Biases, Racialized Organizations, and the Production of Racial/Ethnic Disparities in School Discipline. *American Sociological Review*, 87(6), 1007–1048. https://doi.org/10.1177/00031224221135810

Passaretta, G., and Skopek, J. 2021. "Does Schooling Decrease Socioeconomic Inequality in Early Achievement? A Differential Exposure Approach." *American Sociological Review*. https://doi.org/10.1177/00031224211049188

Petzold, K. 2022. Factorial Survey Experiments in the Sociology of Education. Potentials, Pitfalls, Evaluation. *Swiss Journal of Sociology*, 48(1) 47-76. https://doi.org/10.2478/sjs-2022-0001

Phelps, E. S. 1972. The statistical theory of racism and sexism. *The American Economic Review*, 62(4), 659-661.

Pit-ten Cate, M., I., and Glock, S. 2019. Teachers' Implicit Attitudes Toward Students from Different Social Groups: A Meta-Analysis. *Frontiers in Psychology*, 10, 491099. https://doi.org/10.3389/fpsyg.2019.02832

Polavieja, J. 2023. The Name of the Beast: Hiring discrimination against Spanish-Castilian applicants in Catalonia during the 'Procés'. Presentation at the V Annual Conference of Experimental Sociology, CSIC, Madrid.

Quinn, D. M. 2020. Experimental evidence on teachers' racial bias in student evaluation: The role of grading scales. *Educational Evaluation and Policy Analysis*, 42(3), 375-392.

Ridgeway Cecilia L. 2014. "Why Status Matters for Inequality." *American Sociological Review* 79(1):1–16.

Roscigno, V. J., and Ainsworth-Darnell, J. W. 1999. Race, cultural capital, and educational resources: Persistent inequalities and achievement returns. *Sociology of Education*, 158-178.

Salza, G. 2022. Equally performing, unfairly evaluated: The social determinants of grade repetition in Italian high schools. *Research in Social Stratification and Mobility*. 77.

Schonlau, Matthias, Arthur Van Soest, Arie Kapteyn, and Mick Couper. 2009. "Selection Bias in Web Surveys and the Use of Propensity Scores." *Sociological Methods and Research* 37(3):291-318.

Schuessler, J., and Freitag, M. 2020. Power Analysis for Conjoint Experiments. https://doi.org/10.31235/osf.io/9yuhp

Schuessler, J., and Sønderskov, K. M. 2023,. Compensating Discrimination: Behavioral Evidence from Danish School Registers. https://doi.org/10.31235/osf.io/5zm87

Skopek, J., and Passaretta, G. 2021. Socioeconomic Inequality in Children's Achievement from Infancy to Adolescence: The Case of Germany. *Social Forces*, *100*(1), 86-112.

Small, M. L., Harding, D. and Lamont, M. 2010. 'Reconsidering Culture and Poverty', *ANNALS*, 629, 1 –22.

Spinath, B., and Spinath, F. M. 2005. Development of self-perceived ability in elementary school: the role of parents' perceptions, teacher evaluations, and intelligence. *Cognitive Development*, 20(2), 190–204.

Sprietsma, M. 2013. "Discrimination in grading: Experimental evidence from primary school teachers." *Empirical Economics* 45(1):523-538.

Starck, J. G., Riddle, T., Sinclair, S., and Warikoo, N. 2020. Teachers Are People Too: Examining the Racial Bias of Teachers Compared to Other American Adults. *Educational Researcher*. https://doi.org/10.3102/0013189X20912758

Stefanelli, A., and Lukac, M. 2020. Subjects, Trials, and Levels: Statistical Power in Conjoint Experiments. https://doi.org/10.31235/osf.io/spkcy

Südkamp, A., Kaiser, J., and Möller, J. 2012. Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology*, 104(3), 743-762.

Sullivan, A. 2001. Cultural capital and educational attainment. *Sociology*, 35(4), 893-912.

Sullivan, A. 2002. Bourdieu and Education: How Useful is Bourdieu's Theory for Researchers? *Netherlands Journal of Social Sciences*, 38(2): 144-166

Tajfel, H. and Turner, J.C. 1986. The Social Identity Theory of Intergroup Behavior. In: Worchel, S. and Austin, W.G., Eds., *Psychology of Intergroup Relation*, Hall Publishers, Chicago, 7-24.

Tilcsik, A. 2020. Statistical Discrimination and the Rationalization of Stereotypes. *American Sociological Review*. https://doi.org/10.1177/0003122420969399

Timmermans, A. C., de Boer, H., Amsing, H. T. A., and van der Werf, M. P. C. 2018. Track recommendation bias: Gender, migration background and SES bias over a 20-year period in the Dutch context. *British Educational Research Journal*, 44(5), 847-874

Timmermans, A. C., Kuyper, H., and Werf, G. 2015. Accurate, inaccurate, or biased teacher expectations: Do Dutch teachers differ in their expectations at the end of primary education? *British Journal of Educational Psychology*, 85(4), 459–478.

Tobisch, A., and Dresel, M. 2017. "Negatively or positively biased? Dependencies of teachers' judgments and expectations based on students' ethnic and social backgrounds". *Social Psychology of Education* 20:731–752. https://doi.org/10.1007/s11218-017-9392-z

Triventi, M. 2019. Are Children of Immigrants Graded Less Generously by their Teachers than Natives, and Why? Evidence from Student Population Data in Italy. *International Migration Review*.

Valliant, R., and Dever, J. A. 2018. *Survey weights: a step-by-step guide to calculation* (First edition). Stata Press.

Van de Werfhorst, H. G. 2010. Cultural capital: strengths, weaknesses and two advancements. *British Journal of Sociology of Education*, 31(2), 157-169.

Van de Werfhorst, H. G., and Hofstede, S. 2007. Cultural capital or relative risk aversion? Two mechanisms for educational inequality compared 1. *The British Journal of Sociology*, 58(3), 391-415.

van Huizen, T., Jacobs, M. and Oosterveen, M. 2024. "Teacher bias or measurement error?" arXiv:2401.04200 [econ.EM]. https://doi.org/10.48550/arXiv.2401.04200

Wenz, S. E., and Hoenig, K. 2020. Ethnic and Social Class Discrimination in Education: Experimental Evidence from Germany. *Research in Social Stratification and Mobility*. https://doi.org/10.1016/j.rssm.2019.100461

Xu, J., and Gong, J. 2017. Statistical Discrimination, Taste-based Bias, and Cognitive Bias - Analyzing Grading Bias Caused by Handwriting Quality in a Randomized Control Trial.

Xu, J., and Hampden-Thompson, G. 2012. Cultural reproduction, cultural mobility, cultural resources, or trivial effect? A comparative approach to cultural capital and educational performance. *Comparative Education Review*, 56(1), 98-124.

Yamamoto, Y., and Brinton, M. C. 2010. Cultural capital in East Asian educational systems: The case of Japan. *Sociology of Education*, 83(1), 67-83.

Zanga, G., and De Gioannis, E. 2023. Discrimination in grading: A scoping review of studies on teachers' discrimination in school. *Studies in Educational Evaluation*, *78*, 101284. https://doi.org/10.1016/j.stueduc.2023.101284

7. Annexes

A.1. Experimental Set Up: Pre-tests, Pre-registration, and Timeline

Table A.1. displays the experimental design and fieldwork implementation stages. Before data collection and analysis started, we pre-registered a pre-analysis plan on the *Open Science Foundation* and EGAP Registry on March 31st, 2023; all replication files are available on the corresponding author's *GitHub* account. We applied three pre-tests, reaching 603 observations among in-service (n=503) and pre-service teachers (n=100) to externally validate the relevant features of the essay and the cultural capital instruments. For the pre-test applied to in-service teachers, we contacted all public and private elementary schools in the Spanish regions of Madrid and Andalusia, the two most populated non-bilingual Spanish regions. We used administrative databases of schools' contact emails as a sampling frame (N=3,865 schools). We asked the receiver to forward the invitation email containing the link to the online questionnaire to all elementary education teachers at each school. For the pre-test applied to pre-service teachers (a complete pilot of the final experiment), we contacted one Faculty of Education and asked the Faculty Dean to forward the online questionnaire to all students enrolled in the BA in Primary Education (n=100; 9.4% response rate). Drawing on these pre-tests, in the pre-analysis plan, we defined the study background and objectives, the research hypotheses, and the study methodological design–including methods, measurements, models, power analysis, sampling, and data collection protocols before conducting the fieldwork and data analysis from April 11th to June 5th, 2023, which was discontinued after reaching the minimum projected sample size to detect powered effects.

Table A.1. *Experiment Timeline*

Experiment Phase	2022			2023		
	May-August	September-October	November-December	January-March	April-June	July-December
Research design and survey tools						
Ethics Board review						
Pre-tests and pre-registration						
Data collection						
Analysis and article writing						

A.2. Data Collection Protocols and Ethics

Table A.2. below summarizes the sampled institutions' (see article's section 3.2. for sampling details) population (N), number of participants in our study (n), and response rates. We followed a standardised protocol to contact universities and students. We could not approach our target population directly due to the need to preserve participants' privacy and personal data. To contact faculties of education, which constituted the sampling unit, the first point of contact was the Dean or the Faculty or Academic Secretary. A standardised email was sent to each faculty/university, asking them to get involved in the study. Participation entailed forwarding the invitation to all students enrolled in any grade of the BA or double BA in Primary Education. The invitation e-mail was written in neutral language, not revealing the true scope of the study, and included a link to the experimental survey that respected anonymity. The email emphasised the study's respect for privacy and data protection through informed consent and debriefing, as well as the approval of the study by the European Commission's Joint Research Centre's ethics committee in compliance with European legal standards (clearance received on October 10th, 2023). Additionally, the email asked for the number of enrolled students in their mailing list to estimate response rates accurately.

In the standardised email containing the study invitation addressed to our final target sample, the students, we highlighted monetary incentives for participation: a gift card lottery with two large prizes of 200 euros each and 40 smaller-sized prizes of 50 euros each. Monetary incentives likely incentivized the participation of negatively selected students who otherwise would not have participated in the study. The email also stressed the importance of paying attention, not replying randomly or too fast, and completing the entire survey to be eligible for participation in the gift lottery. The study was implemented using questionnaires and computer-based vignettes randomised on *Qualtrics* software. Most participants accessed the study through an email link on smartphones (for which an ad-hoc adaptation granting legibility was made) or personal computers.

The online questionnaire (median response time = 8.2 minutes) is structured around six screens with the following items and order: Screen 0. Introduction and informed consent; Screen 1. Student's file: Table with student characteristics; Screen 2. Student's essay and first outcome of interest (essay grade); Screen 3. Table with student characteristics and second and third outcomes of interest (expectations about grade retention and continuation in the academic track in high school); Screen 4. Question on respondent's perception of student's parental support (potential mechanism for outcomes 2 and 3); Screen 5. Manipulation checks to assess if respondents correctly remember the levels of the factors; Screen 6. A short questionnaire on respondents' socio-demographic characteristics and attitudes towards educational inequality.

Table A.2. *Population, sample, and response rates*

Selection Order	University / Faculty Anonymized ID	Estimated N (1)	Reported N (2)	Admin. N (3)	Experiment n (4)	Response Rate (4/2)
		Public Institutions				
1	#1 (R) (D)	1,380	1,474	1,475	218	14.79%
2	#2 (D)	2,282	2,290	2,310	57	2.49%
3	#3 (D)	2,019	1,974	1,991	44	2.23%
4	#4 (D)	1,494	1,958	1,456	13	0.66%
5	#5 (D)	1,319	2,287	1,286	80	3.50%
6	#6	1,158	1,169	1,169	75	6.42%
7	#7	1,090	1,036	1,036	45	4.34%
8	#8	962	974	974	11	1.13%
9	#9 (D)	903	881	917	46	5.22%
10	#10 (R)	883	871	906	39	4.48%
11	#11 (R) (D)	821	886	886	50	5.64%
12	#12	782	756	760	51	6.75%
13	#13	578	597	596	21	3.52%
14	#14 (D)	519	546	547	57	10.44%
15	#15 (D)	399	1,505	1,507	221	14.68%
		16,589	19,204	17,816	1,028	5.75%
		Private Institutions				
1	#1	4,750	5,941	5,145	462	7.78%
2	#2 (R) (D)	862	698	1,126	146	20.92%
3	#3	1,170	849	1,257	90	10.60%
5	#4 (R)	306	323	324	22	6.81%
		7,088	7,811	7,852	720	11.53%
		Total				
		23,677	27,015	25,668	1,748	6.97%

Notes: (1) Administrative data: 2020-2022 average used for sampling design in 2022; (2) N reported by each university in personal communications in April-June 2023 for the 2022-2023 academic year; (3) Administrative data: 2022-2023 (provisional estimation); (4) Experimental raw sample; (4) Response rates (4/2); R=Closest replacement unit in the sampling frame; D=University including a Double Degree in Primary Education

A.3. Power Analysis

We did a power analysis before data collection and analysis, as pre-registered in the *Open Science Foundation*. The power of the experiment mainly depends on the following factors: (1) the desired power or probability of correctly rejecting the null hypotheses when the true effect ≠ 0: $1-\beta = 0.8$; (2) the desired statistical significance level: $\alpha = 0.05$ (two-tailed t-test); (3) the expected main effect size (β) on target population, which is likely to be small based on previous research: Cohen's D = 0.1-0.2; Average Marginal Component Effect (AMCE) = 0.05-0.1 (dichotomous outcome scale); unstandardized mean difference = 0.2-0.3 (0-10 or 1-10 scale); and (4) the expected sample size. In the pre-analysis plan, we indicated $n \approx 1{,}367$ under a lower-bound response rate at 5% with one vignette by respondent following the sampling design outlined in the article's section 3.2.

Based on the framework by Hainmuller et al. (2014) and as illustrated in Figure A.1. below, we conducted power calculations for the Average Marginal Component Effect (AMCE) using the R tool developed by Freitag and Schuessler (2020) and for an unstandardized regression coefficient using a SAS software tool (Dziak, Collins and Wagner 2013). The parameters are set at one vignette per respondent and a maximum of 2 levels per attribute. Note that for power calculations, the levels of an attribute do matter, but not the number of attributes (see Schuessler and Freitag 2020).

To come up with the bounds on the effect size, we relied on meta-analyses (Schuessler and Freitag, 2020; Stefanelli and Lukac, 2020), previous observational studies as a reasonable upper-bound (i.e., Gortázar, Martínez de Lafuente and Vega-Bayo 2022; Salza 2022), and the experimental study that most closely resembles our design, that by Wenz and Hoenig (2020). They use two outcomes comparable to ours: grading an essay (0-14 scale, later truncated) and expecting the student to succeed at the Gymnasium (from 1, very unlikely, to 5, very likely, collapsed into three categories). For the essay grade outcome, they find a statistically insignificant main effect of SES that is also relatively small, close to null, and in the opposite direction as ours and their hypothesis: −0.07 (SE 0.16). For teachers' expectations, they find that moving from low to high SES has an average marginal effect of 0.11 but fails to reach conventional statistical significance (p=0.134). Furthermore, the sample size of that study is n=237 teachers; it is most likely underpowered, which casts further doubt on the appropriateness of using their effects as a benchmark for our power calculations.

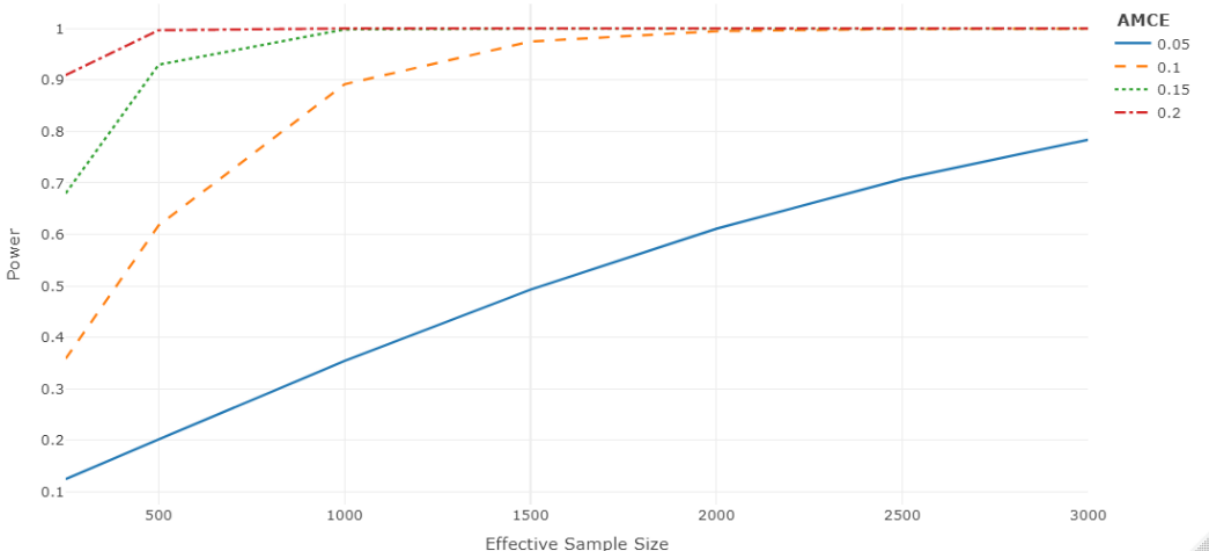
Given that the range of the outcome is different, that they do not find large or statistically significant main SES effects, that their study is most likely underpowered, and that we are not looking at proportions but at mean values (Auspurg and Hinz 2015:33), we find it rather challenging to base calculations on these experimental estimates. Nevertheless, we provided a conservative range of expected effect sizes in the pre-analysis plan according to previous observational and experimental research.

As a conservative best guess, we firstly estimated the minimum detectable effect size with the minimum expected sample size (n=1,367; tasks=1) with power=0.80, two-sided alpha=0.05, and Yσ ≈ 2 at AMCE=0.075 (dichotomous outcome scale), Cohen's D=0.15, or 0.3 raw mean difference (1-10 or 0-10 outcome scale). Secondly, to design the proper sampling procedures to ensure the minimum sample size for the fieldwork, we calculated the minimum sample size necessary to detect the expected main effect with power=80% and two-sided alpha=0.05 at n ≥ 1,398 for an AMCE=0.075 (dichotomous outcome scale), Cohen's D=0.15, or 0.3 raw mean difference (1-10 or 0-10 outcome scale).

In the final experiment, we reached a larger analytical sample (n=1,717; response rate=7%) than estimated in the pre-registered power analysis (n=1,398; lower-bound response rate = 5%), but the effect sizes were also slightly smaller than expected in the pre-analysis plan at, on average, Cohen's D=0.1 or 0.2 raw mean difference (1-10 or 0-10 outcome scale). Thus, we (re)estimated the minimum detectable effect sizes with our final analytical sample (n=1,717) with power=0.8, two-sided alpha=0.05 and the observed SD of our three outcome variables at β = 0.133 (Yσ = 1.96) for essay grading, β = 0.199 (Yσ = 2.95) grade retention recommendations, and β = 0.146 (Yσ = 2.16) for expectations about continuation into the upper-secondary academic track. According to the actual effect sizes of the main models estimated (see M2 in Table A.6. below), some estimations below these thresholds, especially for the outcome on expectations about grade retention (i.e., gender and ethnic-origin coefficients), might be underpowered. Still, looking at a sample of n=1,717, our final analytical sample is a significant improvement from any factorial survey experiment on teachers' bias available so far (Stefanelli and Lukac 2020).

Finally, we used the *cjpowR* R package from Schuessler and Freitag (2020) to conduct a power analysis for interaction effects. We estimate that to identify an Average Marginal Component Interaction Effect (AMCIE) of 5% (7.5%) for a dichotomous outcome scale between attributes of two levels each, we would need a sample of n≥12,118 (n≥5,550). Thus, given the final/analytical sample we reached in the fieldwork (n=1,717), we cannot generally estimate moderation analyses by interacting different factors with enough statistical precision, except when the magnitude of the interaction effect was considerable.

Figure A.1. *Power analysis: Power by Effective Sample Size and AMCE Size (dichotomous outcome scale)*



A.4. Essay Quality Validation and Implementation

Figure A.2. *Essay Grade Distribution by Essay Quality in Pre-Test (in-service teachers, upper-panel) and Experiment (pre-service teachers, bottom-panel)*

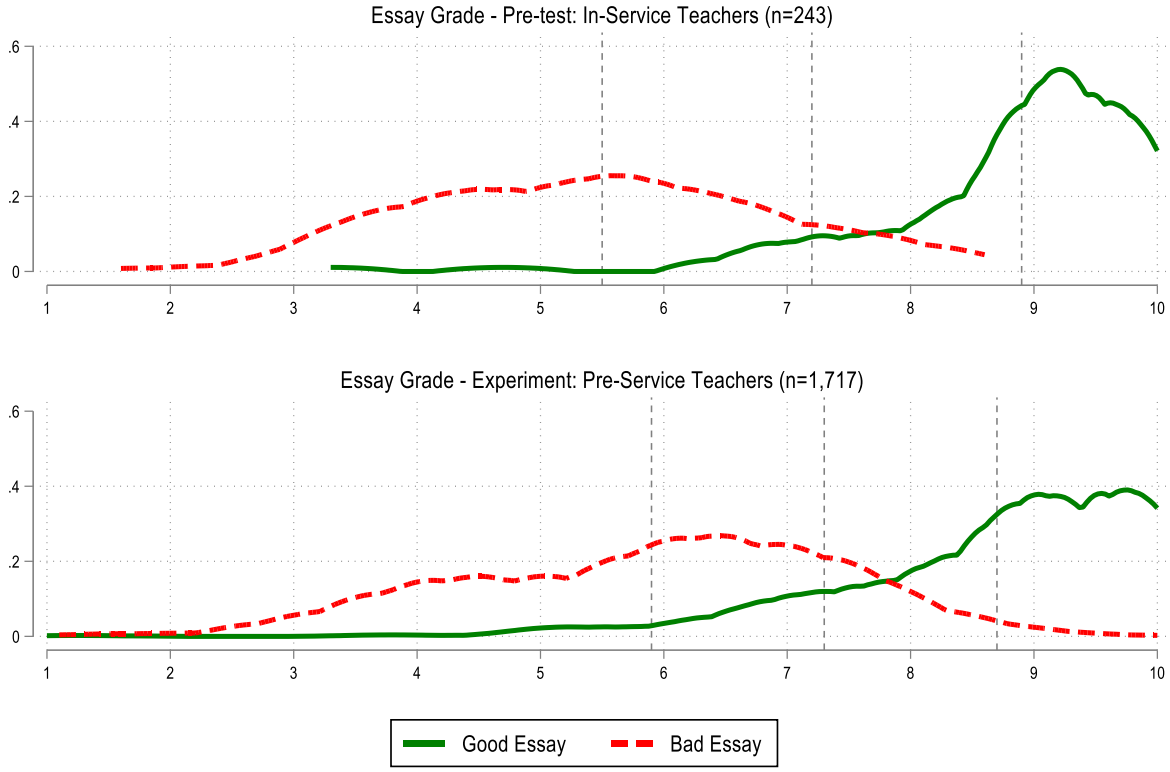


Table A.5. *Essay-screen instructions and essay by objective quality, cultural capital, and parental SES signals (in Spanish)*

A continuación, le presentamos la transcripción de una redacción elaborada por [(Student's) Name Surname(s)], estudiante de 6º de Educación Primaria que le presentamos en la ficha anterior. Por favor, lea el texto con atención. Después le pediremos que evalúe la redacción según criterios de estructura sintáctica, ortografía, vocabulario y creatividad:

1. High Quality Essay (295 words): [**low** / **high** SES; **low** / **high** cultural capital]

*Mi paisaje preferido son los alrededores de un pueblo pequeño que hay no muy lejos de donde vivo. A mi familia y a mí nos encanta pasar tiempo en la naturaleza, todos nos divertimos y mi padre puede desconectar [**de pintar casas en el trabajo** / **del trabajo en la notaría**]. Cuando sales del pueblo puedes disfrutar de paisajes llenos de robles, fresnos y encinas. En algunos prados hay burros que salen a recibirte a los caminos para ver si tienes alguna zanahoria que darles.*

En verano el campo se vuelve amarillo y se llena de cebadillas que se te pegan a los calcetines. En otoño se les caen las hojas a los fresnos y a los robles y la hierba recupera el color verde que la caracteriza. Y llega el invierno, que es la época del fuego; se encienden las chimeneas y se queman las ramas de la poda del verano. Por último, la primavera. Todo se llena de color, a los fresnos les rebrotan las hojas y comienzan a dar sombra y, más tarde, a medida que avanza el calor, los prados se llenan de cardos de todo tipo.

*En el pueblo hay casas muy distintas entre sí, de todos los estilos, gustos y colores posibles. La temperatura es muy variable dependiendo de las estaciones del año; en invierno hace mucho frío y en verano demasiado calor [**, casi como el que pasan en La isla de las tentaciones, que veo en casa en la televisión.** /**. En todas las estaciones los colores me recuerdan a los cuadros impresionistas de Monet que vi en el museo con mi familia.**] Es un pueblo con muchas cuestas; cada vez que paseo por allí acabo casi sin resuello.*

Por la noche se puede oír a las cigarras llamándose unas a otras, a las ranas croando a voz en grito, a las vacas mugiendo, o a los burros rebuznando, ansiosos por comer. La pena es que los humanos estamos acabando con el paisaje y lo vamos a convertir en urbanizaciones y centros comerciales, hasta que hayamos construido hasta en la luna.

0. Low Quality Essay (278 words): [**low** / **high** SES; **low** / **high** cultural capital]

*Mi paisaje preferido es el campo fuera de un pueblecito pequeño al lado de casa. A mi familia y a mi nos encanta pasar tiempo en la naturaleza, todos nos divertimos y mi padre puede desconectar [**de pintar casas en el trabajo** / **del trabajo en la notaría**]. Cuando salgo de el pueblo hay paisajes con un montón de arboles. Los burros salen detrás tuya a los caminos para que les dieras alguna zanaoria. En verano el campo se pone todo amarillo y hay pinchos que se pega a los calcetines y luego en otoño se le cae las hojas a los arboles y ya todo se pone mas verde. Luego llega el invierno que es cuando hace un montón de frio y se enciende las chimeneas y se hace fogatas para quemar las ramas que an cortado en verano. Luego depués llega la primavera y todo se llena de colores, los arboles empiezan a tener ojas otra vez y dar sonbra y ya cuando hace calor en los prados salen matojos que pinchan.*

*Después en el pueblo hay muchas casas cada una distinta, la temperatura cambia mucho en las estaciones en invierno hace mucho frio y en verano hace mucho calor [**, casi como el que pasan en La isla de las tentaciones, que veo en casa en la televisión.** /**. En todas las estaciones los colores me recuerdan a los cuadros impresionistas de Monet que vi en el museo con mi familia.**] Es un pueblo con muchas cuestas enpinadas y cuando paso por alli acabo con los pies echos polvo y me duele la barriga. Despues por las noches se puede oir las chicharras cantando a tope. Tambien a las vacas mujiendo que parece que dicen venir todas que aqui hay mas hierba o a los burros rebufnando que tenian mucha hambre. La cosa es que los hombres nos estamos cargando el campo y lo vamos a hacer todo urbanizaciones y tiendas asta que pongamos casas hasta en la luna.*

A.5. Cultural Capital: Signal and Instrument Validation

Table A.6. *Low / High cultural capital signals embedded in the essay (in Spanish)*

High Quality Essay

En el pueblo hay casas muy distintas entre sí, de todos los estilos, gustos y colores posibles. La temperatura es muy variable dependiendo de las estaciones del año; en invierno hace mucho frío y en verano demasiado calor [, casi como el que pasan en La isla de las tentaciones, que veo en casa en la televisión. /. En todas las estaciones los colores me recuerdan a los cuadros impresionistas de Monet que vi en el museo con mi familia.]

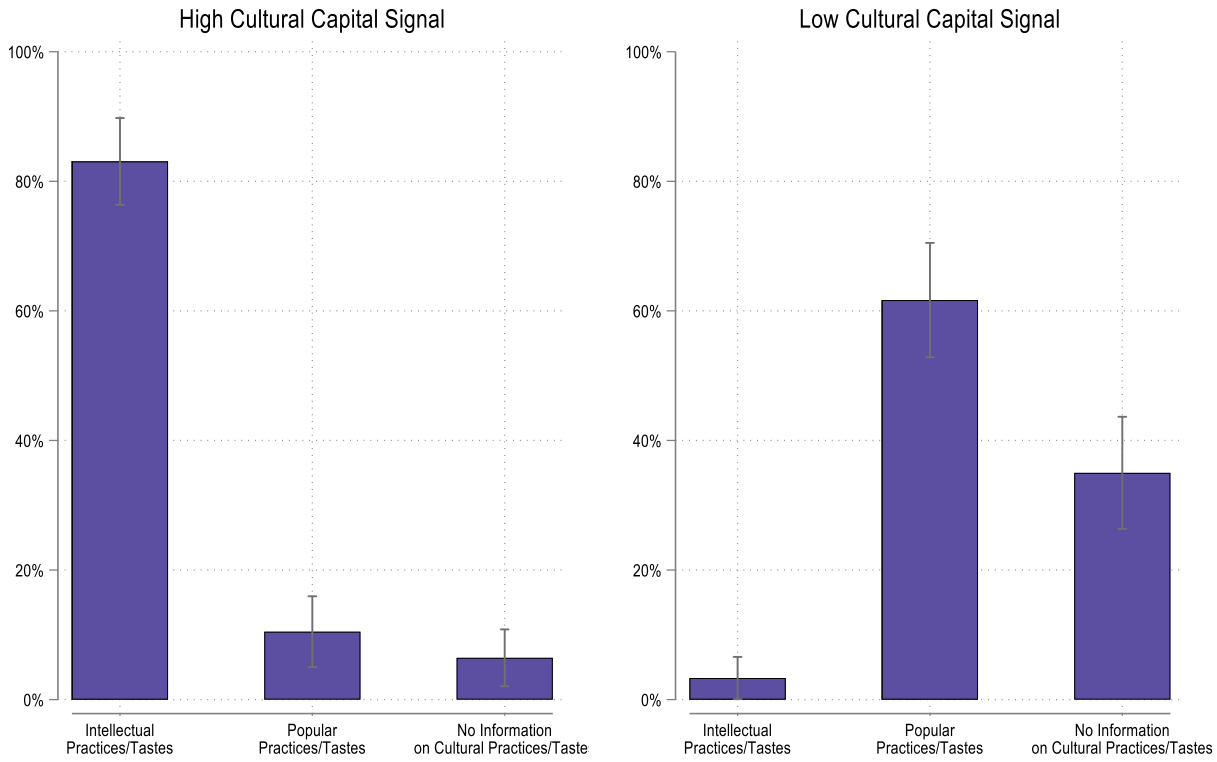
Low Quality Essay

Después en el pueblo hay muchas casas cada una distinta, la temperatura cambia mucho en las estaciones en invierno hace mucho frio y en verano hace mucho calor [, casi como el que pasan en La isla de las tentaciones, que veo en casa en la televisión. /. En todas las estaciones los colores me recuerdan a los cuadros impresionistas de Monet que vi en el museo con mi familia.]

Cultural capital is expressed in three dimensions (Sullivan 2002): (1) embodied through socialization or concerted cultivation (i.e., habitus); (2) objectivized in material cultural resources: books, pieces of art, musical instruments; and (3) institutionalized or formal: certified educational credentials. Previous research examined the following dimensions in the transmission of embodied cultural capital between parents and children (Jæger and Breen 2016), which are claimed to influence students' performance and teachers' biases in assessments: highbrow culture and leisure activities (e.g., going to the opera, ballet, theatre, museums), reading habits (e.g., bedtime reading), cultural communication (i.e., teaching children to be analytical, reasoning, and argumentative), and extracurricular activities (e.g., theatre, conservatory, second-language lessons).

To ensure that the embodied cultural capital signals shown in Table A.6. are actually perceived as highbrow or lowbrow culture by respondents, in our pre-test with 243 in-service elementary education teachers we asked participants to evaluate which kind of information about the cultural practices and tastes of the student and their family the abovementioned cultural capital indicators suggested to them: (1) intellectual cultural practices and tastes; (2) popular culture practices and tastes; or (3) no information about the student and family cultural practices and tastes. Overall, the cultural capital indicators correctly signaled the assumed status hierarchy (Jæger, Rasmussen and Holm 2023; Childress et al. 2022; Lizardo and Skiles 2009) since, as shown in Figure A.3. below, over 80% of respondents associated the cultural reference to visiting an art museum and knowing an impressionist painter with intellectual, cultural practices and tastes, while 60% associated watching a reality show TV programme with popular culture practices and tastes. Still, even when about 35% of respondents claimed that the popular culture reference to watching a *trash* TV programme did not convey any information on the cultural practices and tastes of the student and his family, we suspect that a substantial amount of this share might be hiding social desirability bias and avoiding negative labelling since this was asked openly in the pre-test.

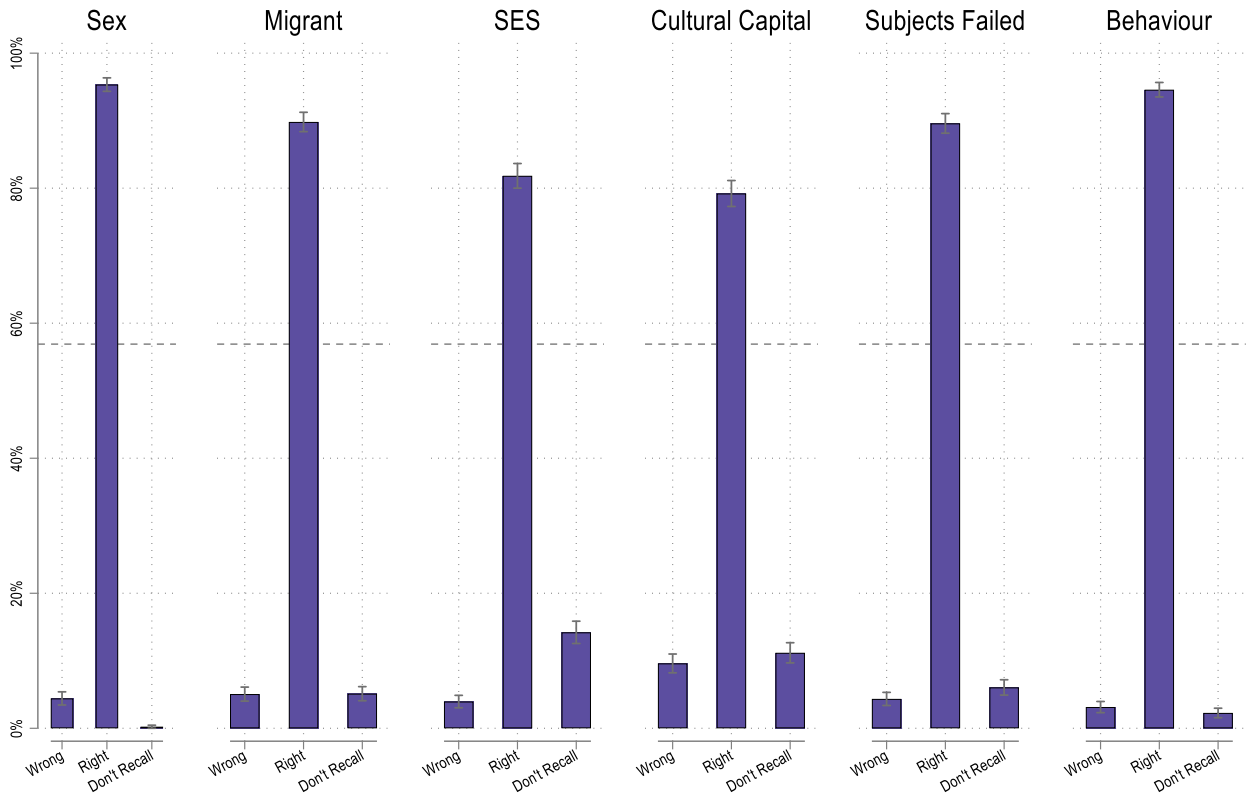
Figure A.3. *Cultural Capital: Pre-test Validation with In-service Teachers (n=243)*



A.6. Manipulation Checks

We included a post-experimental survey module including several questions as manipulation checks to assess the effectiveness of the study's factorial manipulations or randomized treatments. These checks ensure that the signals, such as cultural capital markers (see above), the student's parental SES, ethnic origin, and gender, along with the students' ability-related factors, are working as intended by being correctly recognized and remembered by the respondents. That is key in our design for causally identifying potential biases in respondents' assessments by the randomised treatments while properly controlling for all the relevant confounders. However, not remembering the factors could also be a proxy for not paying enough attention to that information precisely because the participant might not consider it relevant for the required assessment. As shown in Figure A.4., we found that the correct recall of single treatments or factor levels is over 80%, varying from 79% for cultural capital to 95% for gender and behaviour; 57% of respondents correctly recalled all factorial manipulations included. We run robustness checks of all the main analyses on a subsample of respondents correctly recalling all treatments (section A.9.).

Figure A.4. *Manipulation Check by Factor: % Correctly Remembering the Level*



A.7. Vignettes Randomization and Distribution

Figure A.5. *Number of Respondents (n=1,717) by Vignette's Population (n=128)*

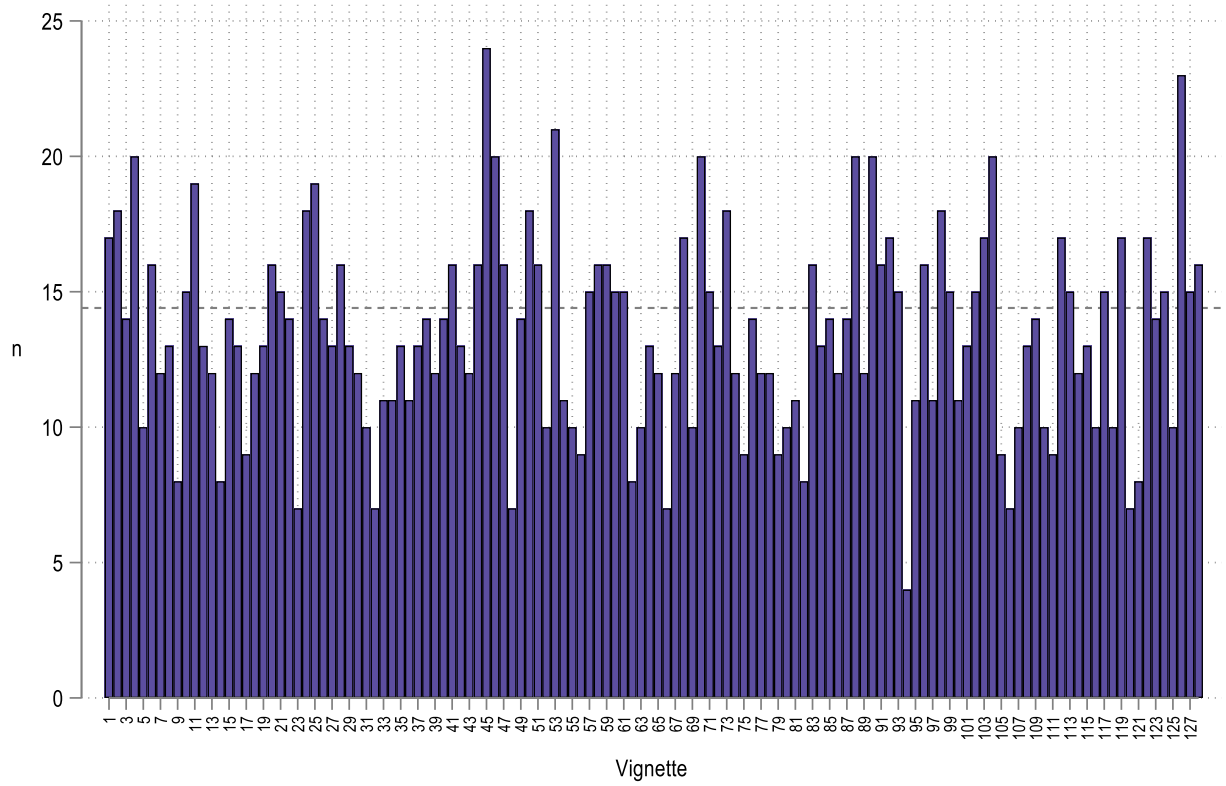
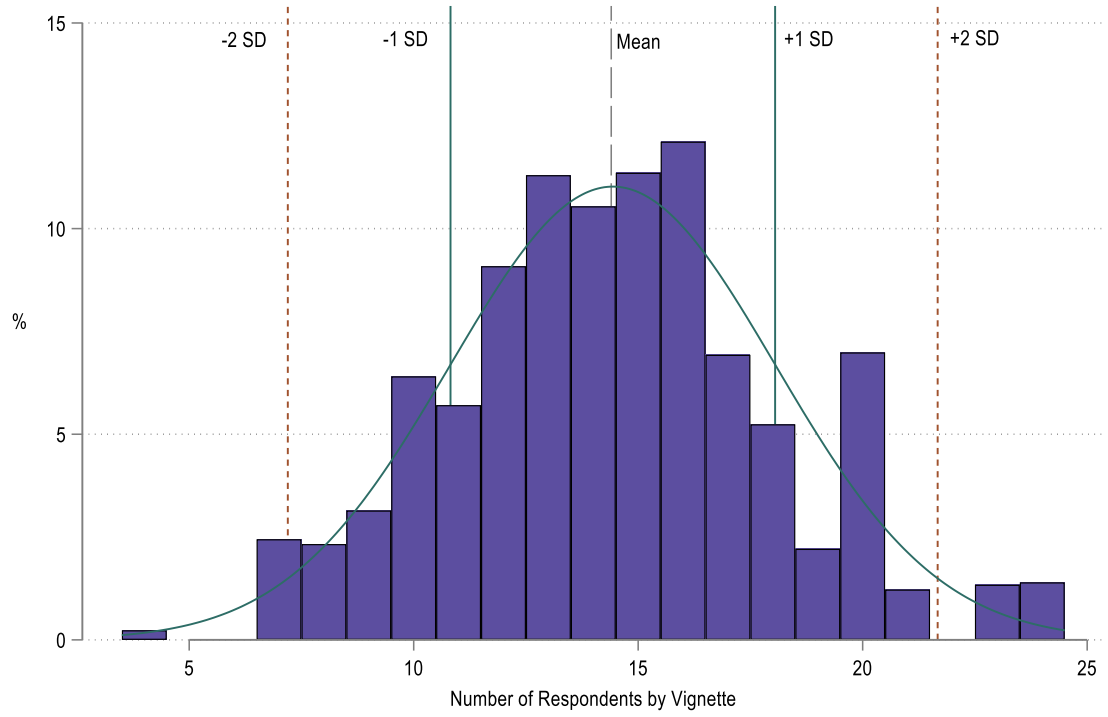


Figure A.6. *Distribution of Number of Respondents (n=1,717) by Vignette (n=128)*



A.8. Main Models' Full Output

Table A.6. *Main OLS models (M1 and M2)*

	Essay Grade (1-10)		Grade Retention Recommendations (0-10)		Academic Track Expectations (0-10)	
	M1	M2	M1	M2	M1	M2
Experimental Factors						
Female	0.124^{+}	0.121^{+}	-0.103	-0.128	0.243**	0.240**
	(0.060)	(0.067)	(0.109)	(0.115)	(0.079)	(0.074)
Spanish Origin	-0.218**	-0.196**	0.170	0.129	0.191*	0.188*
	(0.056)	(0.060)	(0.109)	(0.107)	(0.072)	(0.073)
High-SES	0.0311	0.0335	-0.0198	-0.0266	0.196*	0.199*
	(0.065)	(0.063)	(0.109)	(0.115)	(0.078)	(0.079)
High Cultural Capital	0.204***	0.203***	-0.0911	-0.0859	0.0851	0.0895
	(0.049)	(0.047)	(0.124)	(0.118)	(0.079)	(0.075)
Good Essay	2.836***	2.832***	-2.204***	-2.169***	1.323***	1.313***
	(0.108)	(0.107)	(0.132)	(0.135)	(0.094)	(0.096)
All Subjects Passed	0.282**	0.283**	-1.723***	-1.731***	0.456**	0.465**
	(0.072)	(0.073)	(0.087)	(0.091)	(0.123)	(0.120)
Good Behavior+Effort	0.261**	0.268**	-1.032***	-1.027***	1.207***	1.209***
	(0.080)	(0.078)	(0.103)	(0.095)	(0.100)	(0.097)
Individual-Level Characteristics						
Year of Birth		0.00936		0.00248		0.0127
		(0.006)		(0.009)		(0.008)
Female		0.0136		0.156		-0.0558
		(0.054)		(0.119)		(0.093)
2nd Grade (1st Grade)		0.223*		-0.240		0.0682
		(0.105)		(0.216)		(0.129)
3rd Grade		0.255*		-0.512*		0.170
		(0.104)		(0.186)		(0.141)
4th Grade		0.272**		-0.514*		0.0587
		(0.076)		(0.197)		(0.157)
5th Grade		0.387*		-0.552^{+}		0.0329
		(0.159)		(0.268)		(0.314)
Graduated		0.0941		-0.946^{+}		-0.0329
		(0.264)		(0.458)		(0.269)
Grade Retention		-0.0724		-0.00791		0.129
		(0.080)		(0.126)		(0.141)
Low-SES		-0.111		0.190*		-0.0594
		(0.076)		(0.087)		(0.080)
Foreign-Born		-0.00923		0.126		0.308
		(0.173)		(0.343)		(0.253)
Foreign-Born Parents		0.169		-0.337		0.128
		(0.125)		(0.244)		(0.323)
Institution FE		✓		✓		✓
Observations	1,717	1,717	1,717	1,717	1,717	1,717
Adjusted R^2	0.518	0.522	0.245	0.254	0.180	0.186

Notes: Clustered standard errors by institutions in parentheses; $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

A.9. Robustness Checks

Table A.7. *Manipulation check: main model M2 and M2 among the subsample correctly recalling all signals (M2 | Signals)*

	Essay Grade (1-10)		Grade Retention Recommendations (0-10)		Academic Track Expectations (0-10)	
	M2	M2 | Signals	M2	M2 | Signals	M2	M2 | Signals
Female	0.121^{+}	0.144^{+}	-0.128	-0.194	0.240**	0.366**
	(0.067)	(0.069)	(0.115)	(0.149)	(0.074)	(0.123)
Native Origin	-0.196**	-0.209*	0.129	0.0591	0.188*	0.109
	(0.060)	(0.087)	(0.107)	(0.167)	(0.073)	(0.107)
High-SES	0.0335	0.0919	-0.0266	0.00618	0.199*	0.214*
	(0.063)	(0.069)	(0.115)	(0.176)	(0.079)	(0.098)
High Cultural Capital	0.203***	0.299***	-0.0859	-0.000210	0.0895	0.183
	(0.047)	(0.072)	(0.118)	(0.122)	(0.075)	(0.116)
Good Essay	2.832***	2.999***	-2.169***	-2.374***	1.313***	1.487***
	(0.107)	(0.095)	(0.135)	(0.127)	(0.096)	(0.119)
All Subjects Passed	0.283**	0.267*	-1.731***	-1.984***	0.465**	0.518**
	(0.073)	(0.125)	(0.091)	(0.175)	(0.120)	(0.167)
Good Behavior+Effort	0.268**	0.232*	-1.027***	-1.048***	1.209***	1.221***
	(0.078)	(0.093)	(0.095)	(0.151)	(0.097)	(0.210)
Institution FE	✓	✓	✓	✓	✓	✓
Individual Controls	✓	✓	✓	✓	✓	✓
Observations	1,717	977	1,717	977	1,717	977
Adjusted R^2	0.522	0.555	0.254	0.314	0.186	0.226

Notes: Clustered standard errors by institutions in parentheses, individual-level controls: year of birth, gender, country of birth, parental country of birth, parental highest education, BA Degree enrollment grade, grade retention in primary and/or secondary school. Two-tailed t-tests: $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A.8. *Main models without and with weighting by population sociodemographics*

	Essay Grade (1-10)		Grade Retention Recommendations (0-10)		Academic Track Expectations (0-10)	
	M2	M2 Weighted	M2	M2 Weighted	M2	M2 Weighted
Experimental Factors						
Female	0.121^{+}	0.165*	-0.128	-0.206^{+}	0.240**	0.232*
	(0.067)	(0.078)	(0.115)	(0.118)	(0.074)	(0.084)
Native Origin	-0.196**	-0.183**	0.129	0.127	0.188*	0.200**
	(0.060)	(0.053)	(0.107)	(0.113)	(0.073)	(0.061)
High-SES	0.0335	0.0659	-0.0266	-0.0336	0.199*	0.214*
	(0.063)	(0.070)	(0.115)	(0.124)	(0.079)	(0.090)
High Cultural Capital	0.203***	0.206**	-0.0859	-0.165	0.0895	0.121
	(0.047)	(0.061)	(0.118)	(0.101)	(0.075)	(0.083)
Good Essay	2.832***	2.784***	-2.169***	-2.091***	1.313***	1.263***
	(0.107)	(0.122)	(0.135)	(0.149)	(0.096)	(0.105)
All Subjects Passed	0.283**	0.279***	-1.731***	-1.748***	0.465**	0.424**
	(0.073)	(0.064)	(0.091)	(0.097)	(0.120)	(0.116)
Good Behavior+Effort	0.268**	0.281**	-1.027***	-0.942***	1.209***	1.132***
	(0.078)	(0.093)	(0.095)	(0.087)	(0.097)	(0.096)
Individual-Level Characteristics						
Year of Birth	0.00936	0.0103	0.00248	0.00572	0.0127	0.00873
	(0.006)	(0.007)	(0.009)	(0.007)	(0.008)	(0.007)
Female	0.0136	-0.00551	0.156	0.174	-0.0558	-0.0159
	(0.054)	(0.053)	(0.119)	(0.127)	(0.093)	(0.097)
2nd Grade (1st)	0.223*	0.225*	-0.240	-0.272	0.0682	0.104
	(0.105)	(0.096)	(0.216)	(0.275)	(0.129)	(0.156)
3rd Grade	0.255*	0.266^{+}	-0.512*	-0.570**	0.170	0.185
	(0.104)	(0.133)	(0.186)	(0.186)	(0.141)	(0.161)
4th Grade	0.272**	0.296*	-0.514*	-0.634**	0.0587	0.0741
	(0.076)	(0.107)	(0.197)	(0.166)	(0.157)	(0.178)
5th Grade	0.387*	0.452^{+}	-0.552^{+}	-0.639*	0.0329	-0.0246
	(0.159)	(0.237)	(0.268)	(0.224)	(0.314)	(0.280)
Graduated	0.0941	0.0551	-0.946^{+}	-0.912^{+}	-0.0329	-0.214
	(0.264)	(0.300)	(0.458)	(0.443)	(0.269)	(0.269)
Grade Retention	-0.0724	-0.0514	-0.00791	-0.0363	0.129	0.175
	(0.080)	(0.112)	(0.126)	(0.176)	(0.141)	(0.133)
Low-SES	-0.111	-0.134	0.190*	0.160^{+}	-0.0594	-0.0205
	(0.076)	(0.080)	(0.087)	(0.082)	(0.080)	(0.085)
Foreign-Born	-0.00923	-0.143	0.126	0.117	0.308	0.163
	(0.173)	(0.182)	(0.343)	(0.369)	(0.253)	(0.317)
Foreign-Born Parents	0.169	0.341*	-0.337	-0.406	0.128	0.161
	(0.125)	(0.132)	(0.244)	(0.321)	(0.323)	(0.337)
Institution FE	✓	✓	✓	✓	✓	✓
Observations	1,717	1,717	1,717	1,717	1,717	1,717
Adjusted R^2	0.522	0.515	0.254	0.251	0.186	0.171

Notes: Clustered standard errors by institutions in parentheses; $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A.9. *Main OLS models and LPM with dummy outcomes (below/above median)*

	Essay Grade		Grade Retention Recommendations		Academic Track Expectations	
	OLS (1-10)	LPM (0-1)	OLS (1-10)	LPM (0-1)	OLS (1-10)	LPM (0-1)
Female	0.121^{+}	0.0151	-0.128	-0.0291	0.240**	0.0528***
	(0.067)	(0.016)	(0.115)	(0.017)	(0.074)	(0.013)
Native Origin	-0.196**	-0.0332*	0.129	0.0195	0.188*	0.0315
	(0.060)	(0.014)	(0.107)	(0.022)	(0.073)	(0.020)
High-SES	0.0335	0.0119	-0.0266	-0.00260	0.199*	0.0237
	(0.063)	(0.011)	(0.115)	(0.021)	(0.079)	(0.016)
High Cultural Capital	0.203***	0.0372**	-0.0859	-0.0142	0.0895	0.0170
	(0.047)	(0.011)	(0.118)	(0.014)	(0.075)	(0.015)
Good Essay	2.832***	0.728***	-2.169***	-0.337***	1.313***	0.279***
	(0.107)	(0.025)	(0.135)	(0.019)	(0.096)	(0.025)
All Subjects Passed	0.283**	0.0421*	-1.731***	-0.304***	0.465**	0.122***
	(0.073)	(0.018)	(0.091)	(0.013)	(0.120)	(0.025)
Good Behavior+Effort	0.268**	0.0193	-1.027***	-0.152***	1.209***	0.274***
	(0.078)	(0.017)	(0.095)	(0.015)	(0.097)	(0.016)
Institution FE	✓	✓	✓	✓	✓	✓
Individual Controls	✓	✓	✓	✓	✓	✓
Observations	1717	1717	1717	1717	1717	1717
Adjusted R^2	0.522	0.524	0.254	0.227	0.186	0.167

Notes: Clustered standard errors by institutions in parentheses, individual-level controls: year of birth, gender, country of birth, parental country of birth, parental highest education, BA Degree enrollment grade, grade retention in primary and/or secondary school. Two-tailed t-tests: $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure A.7. *OLS-M2 on Essay Grading by Objective Essay Quality (95% CI)*

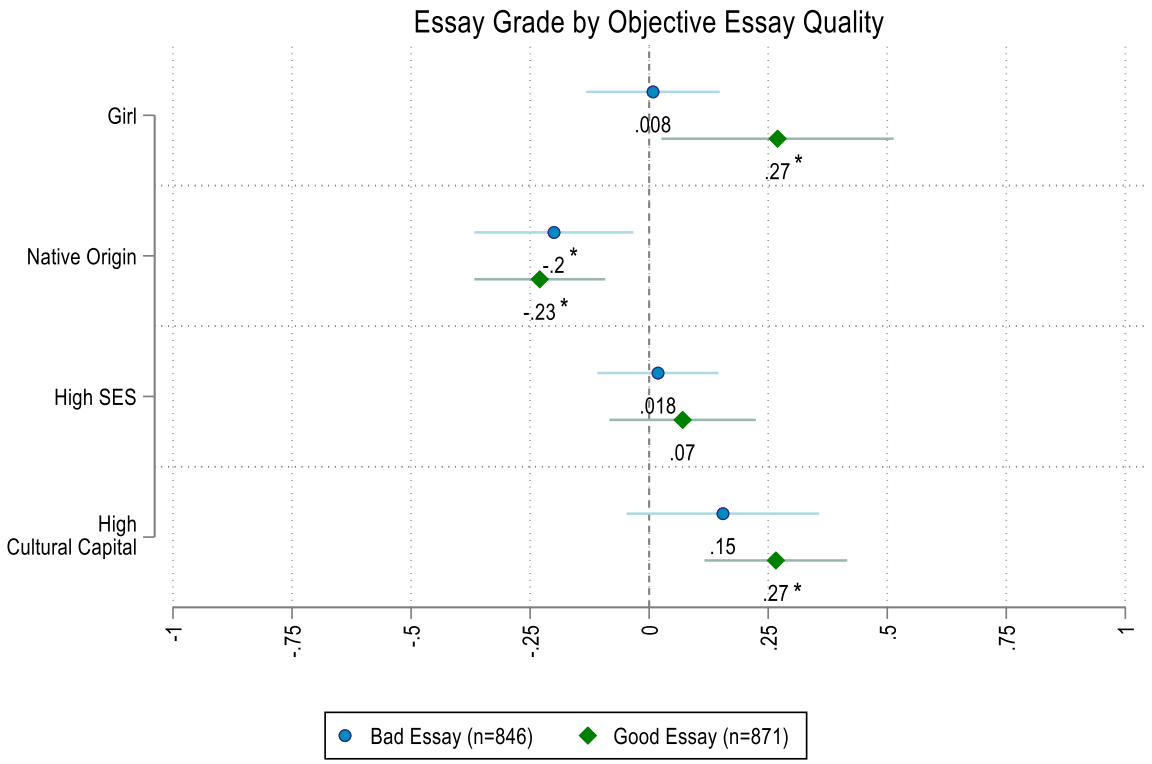
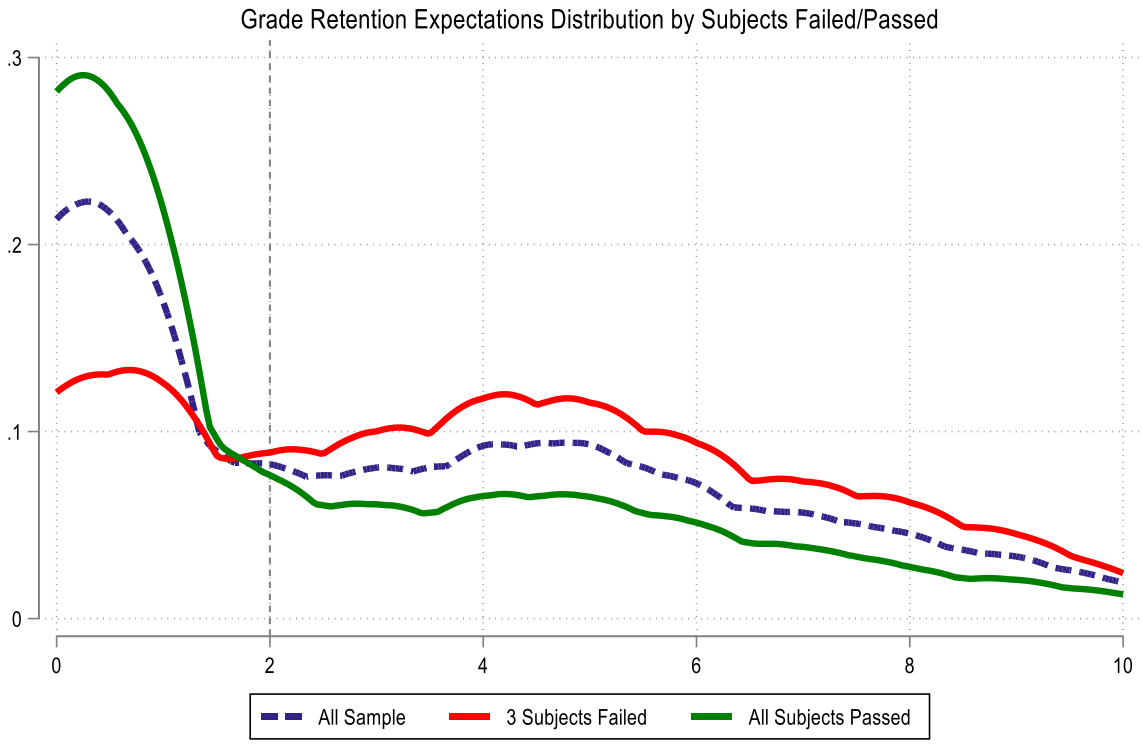
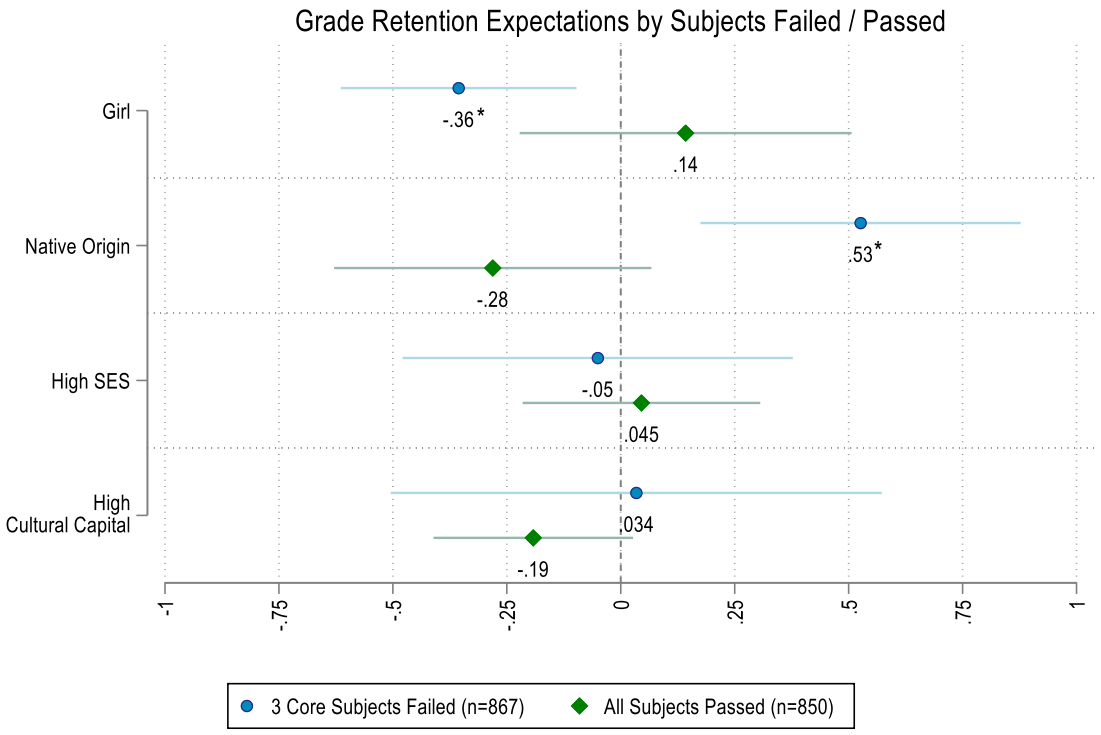


Figure A.8. *Kernel Density of Grade Retention Recommendations by Subjects Failed/Passed*



Notes: Median all sample = 2

Figure A.9. *OLS-M2 on Grade Retention Recommendations by Subjects Failed (95% CI)*



A.10. Mechanisms

Table A.10. *Mechanisms: Ascriptive Factors on Parental Support*

Parental Support (0-10)	
	M2
Female	0.0526
	(0.109)
Spanish Origin	-0.0852
	(0.116)
High-SES	0.142
	(0.122)
High Cultural Capital	0.497***
	(0.087)
Good Essay	1.115***
	(0.088)
All Subjects Passed	0.621***
	(0.104)
Good Behavior + Effort	2.180***
	(0.094)
Institution FE	✓
Individual Controls	✓
Observations	1,717
Adjusted R^2	0.240

Notes: Clustered standard errors by institutions in parentheses, individual-level controls: year of birth, gender, country of birth, parental country of birth, parental highest education, BA Degree enrollment grade, grade retention in primary and/or secondary school. Two-tailed t-tests: $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A.11. *Mechanisms: KHB Linear Models on Confounding / Mediation of Parental Support on Outcomes by Ascriptive Factors*

	Essay Grade (1-10)	Grade Retention Expectations (0-10)	Academic Track Expectations (0-10)
		Gender	
Reduced	0.121*	-0.128	0.240***
	(0.0557)	(0.120)	(0.0586)
Full	0.113*	-0.113	0.223***
	(0.0555)	(0.120)	(0.0586)
Difference	0.00777	-0.0154	0.0169
	(0.0423)	(0.0838)	(0.0919)
		Ethnic Origin	
Reduced	-0.196***	0.129	0.188**
	(0.0515)	(0.0928)	(0.0612)
Full	-0.184***	0.105	0.215***
	(0.0509)	(0.0933)	(0.0609)
Difference	-0.0126	0.0249	-0.0273
	(0.0423)	(0.0838)	(0.0919)
		SES	
Reduced	0.0335	-0.0266	0.199**
	(0.0718)	(0.111)	(0.0673)
Full	0.0125	0.0150	0.153*
	(0.0731)	(0.110)	(0.0677)
Difference	0.0210	-0.0416	0.0456
	(0.0424)	(0.0839)	(0.0919)
		Cultural Capital	
Reduced	0.203***	-0.0859	0.0895
	(0.0462)	(0.115)	(0.0604)
Full	0.130**	0.0593	-0.0697
	(0.0460)	(0.111)	(0.0640)
Difference	**0.0733+**	-0.145+	0.159+
	(0.0431)	(0.0851)	(0.0924)
Mediation / Confound % by Parental Support	**36.13**	169.0	177.9
Observations	1,717	1,717	1,717

Notes: Reduced: M2; Full: Control for parental support; Diff: Factors' coefficients reduction after controlling for parental support. Clustered standard errors by institutions in parentheses. Two-tailed t-tests: + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Individual-level controls: year of birth, gender, country of birth, parental country of birth, parental highest education, BA Degree enrollment grade, grade retention in primary and/or secondary school.

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct centres. You can find the address of the centre nearest you online (european-union.europa.eu/contact-eu/meet-us_en).

On the phone or in writing

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

— by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

— at the following standard number: +32 22999696,

— via the following form: european-union.europa.eu/contact-eu/write-us_en.

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website (european-union.europa.eu).

EU publications

You can view or order EU publications at op.europa.eu/en/publications. Multiple copies of free publications can be obtained by contacting Europe Direct or your local documentation centre (european-union.europa.eu/contact-eu/meet-us_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex (eur-lex.europa.eu).

Open data from the EU

Science for policy

The Joint Research Centre (JRC) provides independent, evidence-based knowledge and science, supporting EU policies to positively impact society

EU Science Hub
joint-research-centre.ec.europa.eu

@EU_ScienceHub

EU Science Hub - Joint Research Centre

EU Science, Research and Innovation

EU Science Hub

@eu_science