sociological science

Teacher Bias in Assessments by Student Ascribed Status: A Factorial Experiment on Discrimination in Education

Carlos J. Gil-Hernández,[a] Irene Pañeda-Fernández,[b] Leire Salazar,[c] Jonatan Castaño Muñoz[d]

a) University of Florence; b) WZB Berlin Social Science Center; c) Consejo Superior de Investigaciones Científicas; d) Universidad de Sevilla

Abstract: Teachers are the evaluators of academic merit. Identifying if their assessments are fair or biased by student-ascribed status is critical for equal opportunity but empirically challenging, with mixed previous findings. We test *status characteristics beliefs*, *statistical discrimination*, and *cultural capital theories* with a pre-registered factorial experiment on a large sample of Spanish pre-service teachers ($n = 1,717$). This design causally identifies, net of ability, the impact of student-ascribed characteristics on teacher short- and long-term assessments, improving prior studies' theory testing, confounding, and power. Findings unveil teacher bias in an essay grading task favoring girls and highbrow cultural capital, aligning with status characteristics and cultural capital theories. Results on teachers' long-term expectations indicate 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 short-term evaluations, suggesting compensatory grading. We discuss the complex roots of discrimination in teacher assessments as an educational (in)equality mechanism.

Keywords: Discrimination; Teacher bias; Assessments; Educational inequality; Factorial experiment; Cultural Capital

Replication Package: Data and replication code are publicly accessible at the *GitHub* repository: https://zenodo.org/doi/10.5281/zenodo.12666534. The hypotheses and research design were publicly pre-registered with a pre-analysis plan (PAP) before data collection and analysis at the *Open Science Foundation* repository: https://doi.org/10.17605/OSF.IO/DZB3S.

Citation: J. Gil-Hernández, Carlos, Irene Pañeda-Fernández, Leire Salazar, Jonatan Castaño Muñoz, 2024. "Teacher Bias in Assessments by Student Ascribed Status: A Factorial Experiment on Discrimination in Education" Sociological Science 11: 743-776.

Received: January 6, 2024

Accepted: July 9, 2024

Published: August 27, 2024

Editor(s): Ari Adut, Stephen Vaisey

DOI: 10.15195/v11.a27

STUDENTS' ascribed characteristics persistently shape educational inequalities. Pupils from high socioeconomic status (SES) (Chmielewski 2019), non-migrant backgrounds (Heath and Brinbaum 2007), and girls (DiPrete and Buchmann 2013) systematically excel at school. The role of families and school context (Downey and Condron 2016) has been extensively scrutinized to explain 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 assessments and students' scores in blind standardized tests (Südkamp, Kaiser, and Möller 2012)—a residual approach assumed as evidence of teacher bias.

Teachers are the primary evaluators of academic merit in the educational system (Bourdieu and Passeron 1990). However, identifying their 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 cognition, attitudes and stereotypes about students, potentially leading to discriminatory evaluations (Fazio et al. 2023). Such biased assumptions may breed self-fulfilling prophecies impeding student progress (Carlana 2019) because teachers' inputs are the main signals for families to navigate the educational system (Holm, Hjorth-Trolle, and Jæger 2019). Therefore, understanding if teachers' assessment practices are fair is crucial to ensuring equal educational opportunity.

Net of performance in test scores, observational research identified a residual association, interpreted as teacher bias, between students' ascribed characteristics, grading (Schuessler and Sønderskov 2023), expectations (Timmermans, Kuyper, and van der Werf 2015), and recommendations (Batruch et al. 2023; Salza 2022; Timmermans et al. 2018). On average, students who are girls (Marcenaro-Gutiérrez and Vignoles 2015), from native origin (Kisfalusi, Janky and Takács 2021), high-SES families (Gortázar, Martínez de Lafuente, and Vega-Bayo 2022), or display high cultural capital (Jæger and Møllegaard 2017) tend to be more positively evaluated at school.[1] Likewise, emerging behavioral (Carlana, La Ferrera, and Pinotti 2022) and experimental studies (Zanga and De Gioannis 2023) document that teacher assessments are influenced by student-ascribed features, like gender and ethnicity (Lorenz et al. 2024).

Despite accumulating evidence, the role of discrimination in education has been underexplored compared to the labour market, with few studies applying experimental designs (Batruch et al. 2023; Zanga and De Gioannis 2023). Teacher bias in assessments remains insufficiently understood due to critical methodological flaws: omitted variable bias and measurement error in observational studies (van Huizen, Jacobs, and Oosterveen 2024), weak reliability in cognitive measures of implicit bias (Miles, Charron-Chénier, and Schleifer 2019), as well as underpowered (Schuessler and Freitag 2020) and low-externally valid experiments (Petzold 2022). Furthermore, most previous studies focused on ethnic or gender discrimination, particularly for grading outcomes (Zanga and De Gioannis 2023), leaving SES- and cultural capital-based biases under-researched. Importantly, no previous study has experimentally disentangled 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 assessment biases by several students' ascribed factors, framing our pre-registered hypotheses in an interdisciplinary discrimination framework spanning sociology, psychology, and economics.[2] Although most previous research only tested single discrimination theories (Correll and Benard 2006), we draw on complementary theories of status characteristics beliefs (Ridgeway 2014), implicit bias (Greenwald and Banaji 1995), statistical discrimination (Arrow 1973), and cultural capital (Jæger and Breen 2016). All of them hypothesize negative teacher bias against boys, ethnic minorities, low-SES, and low-brow cultural capital. We test the explanatory power of these theories by comparing three educational outcomes that, given fixed student information, convey different degrees of uncertainty for teacher evaluations in the short- and long-term—essay grading, grade retention recommendations, and track enrollment expectations.

Our pre-registered experimental and sampling design further contributes to four main methodological fronts. First, individual biases by students' backgrounds are hard to capture with observational data due to social desirability and the impossibility of measuring all (un)observable student characteristics—true ability and behavior (Ferman and Fontes 2022). These are generally proxied with low-stakes competence tests subject to measurement error (Südkamp et al. 2012), which, once corrected, might decrease SES and cultural capital discrimination estimates substantially (van Huizen, Jacobs, and Oosterveen 2024; Jæger 2022) or even change the bias direction favouring ethnic minorities (Zhu 2024). To address these issues, we designed a 2^7-factorial experiment with 128 profiles to isolate the causal effect of students' ascribed characteristics on three teacher's assessments. To identify students' ascribed status—SES, migrant background, gender, and cultural capital—we experimentally manipulate three ability dimensions to rule out confounding: language skills, subjects passed/failed, and socio-emotional skills.

Second, laboratory and factorial experiments are often criticized for their lower validity (Krolak-Schwerdt et al. 2017) relative to field experiments or automatic cognition measures like the *Implicit Association Test* (IAT) (Melamed et al. 2019)—even though IATs are not without issues (Mitchell and Tetlock 2017). Thus, to increase our design validity, we randomly assigned different versions of a real task, an essay written by a sixth grader, experimentally manipulating its objective quality and cultural capital (Farkas 2003). We also untangle parental SES from ethnic origin. Students' SES is usually signaled with names and surnames (Wenz and Hoenig 2020), yet participants might not correctly identify SES variation within foreign origin names (Crabtree et al. 2022). Although we signal gender and migrant origin with name and surname, we subtly embed the students' SES (father's occupation) within the essay and a fictitious but realistic student file (contact email). Thus, this study contributes substantially by experimentally disentangling students' ascribed and ability factors using realistic and externally validated instruments.

Third, instead of in-service teachers, we sampled pre-service teachers—students of the Bachelor of Arts (BA) in Primary Education—to identify if they already show assessment biases well before interacting with students or being exposed to schools. 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 might reinforce or mitigate pre-existing 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 in teacher training programmes (Lehmann-Grube, Tobisch, and Dresel 2023).

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. We recruited a sample of 19 public and private Spanish universities by contacting all Primary Education BA students to reach 1,717 valid respondents. This large sample allowed us to identify powered main effects using a pre-registered power plan.

Findings unveil teacher discrimination in essay grading, showing a preference for girls and highbrow cultural capital, aligning with status characteristics beliefs, implicit bias, and cultural capital theories. Regarding teachers' future educational expectations, findings suggest statistical discrimination against boys, migrant origin, and low-SES students. Surprisingly, the ethnic bias shifts from favoring native origin students in teachers' long-term expectations to *compensatory* grading favoring those with a migrant background.

Theoretical Background, Previous Findings, and Hypotheses

This section outlines theories explaining how teachers might generate observed achievement gaps by student-ascribed status. We expand on each theory and how to differentiate between them by focusing on their observable, testable implications while reviewing previous relevant findings.

Implicit Bias and Status Characteristics Beliefs

Psychological *theories of implicit bias* explain how micro-processes subtly generate social inequality (Fazio et al. 2023; Greenwald and Banaji 1995). Implicit cognition (Greenwald and Krieger 2006) is an unconscious process that might lead to discrimination via two interrelated processes: (1) a tendency to like or dislike group members (implicit attitudes) and (2) the association of a group with a positive or negative trait (implicit stereotypes). Studies deploying implicit bias tests in educational contexts identified teachers' negative reactions against immigrants and low-SES students (Carlana et al. 2022; Pit-ten Cate and Glock 2019; Alesina et al. 2018). Results on gender are mixed, with girls perceived as less proficient in science or math, according to the *Gender-Science IAT* (Carlana 2019), while more skillful in language tasks than boys (Glock and Klapproth 2017). Implicit associations do not necessarily align with explicit behavior (Glock and Krolak-Schwerdt 2014) and might remain unconscious until triggered.

Under a sociological lens, implicit biases emerge during early socialization. They are stored in implicit memory as cultural schemata (DiMaggio 1997) because people process information consistent with pre-existing mental structures. *Status Characteristics Theory* (SCT) focuses on *beliefs* about which social groups are more competent or deserving (Berger et al. 1977). Such beliefs naturally emerge in small-group interactions, falling along ascribed groups—ethnicity, gender, and class (Foley 2023)—as long as these convey distinctive *status* and are salient to the task (Ridgeway 2014). In sum, SCT is a *status generalization* theory attributing abilities to individuals based on group characteristics (Correll and Ridgeway 2006:33). Yet individuals may not realize they hold differential competence *expectations*, linking sociological SCT to psychological implicit bias theories (Melamed et al. 2019).

SCTs have rarely been applied to the educational context. Still, similar status generalization processes might arise when teachers evaluate performance (Kisfalusi, Janky, and Takács 2018). The *Double Standards Theory* (DST) (Foschi 2000) posits that *standards* tied to status characteristics might result in differential performance expectations and biased assessments among equally competent students. Due to entrenched status beliefs, lower-status individuals must outperform higher-status peers for equal task competence recognition because high performance would be inconsistent with their bottom status. DST reveals harsher scrutiny for lower-status individuals, favoring lenient judgment for equally competent higher-status counterparts. Accordingly, teachers' implicit beliefs and expectations about competence and deservingness might contribute to reproducing the observed educational gaps by student-ascribed characteristics, net of their objective ability. These expectations alone might generate gaps via self-fulfilling prophecies (Merton 1968).

Gender is a powerful status characteristic conveying "cultural expectations for competence", where men are typically assumed to be better than women at most tasks (Correll and Ridgeway 2006). However, in the school context, teachers might form implicit biases and status characteristics beliefs by internalizing stereotypes about girls doing better than boys because, currently, it is the case (DiPrete and Buchmann 2013). Accordingly, teachers generally report girls as more academically competent than boys (Homuth, Thielemann, and Wenz 2023), particularly in language—the task evaluated in this article—compared to scientific or math proficiency (Krkovic et al. 2014). We similarly argue that teachers may internalize negative stereotypes about low-SES and migrant origin individuals, given that these abound in Western countries, including Spain (Cea D'Ancona 2016). Besides, these groups objectively underperform native origin and high-SES students (Skopek and Passaretta 2021) and are perceived accordingly by teachers (Homuth et al. 2023). This internalization process may begin before teachers enter service during their schooling and pre-service training through exposure to ascribed groups as classmates.

In sum, teachers' evaluations are tainted by their tendency to like or dislike particular groups, or their differing expectations, beliefs, and standards about the competence of individuals belonging to an ascribed group. In contrast with *statistical discrimination* perspectives discussed below, teacher bias is fairly stable because, with new individual input, teachers will stick to pre-existing status beliefs. Still, although a single individual interaction is unlikely to change behavior, teachers consistently exposed to counter-stereotypical exchanges might decrease bias (Elwert et al. 2023).

Hypothesis 1 (H1). Implicit bias, beliefs and standards about student status characteristics drive teacher evaluations by over-grading (H1a), recommending less grade retention (H1b) and expressing higher expectations (H1c) for girls (vs boys), natives (vs migrants), and high-SES pupils (vs low-SES).

Statistical Discrimination

Statistical discrimination theories, mainly by economists (Arrow 1998, 1973; Borjas and Goldberg 1978; Phelps 1972), have a crucial distinction with implicit bias or SCT. Rather than resulting from deep-rooted beliefs and expectations, discrimination happens due to a lack of perfect information and diminishes once obtained. According to the original formulation applied to labor markets, under imperfect information about true employees' productivity, the employer's rational action is proxying 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.

Although initially developed to explain hiring discrimination, this theory has recently been applied to the educational context. 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), teachers rely less on them, reducing bias against low-caste students, as information about the testing instrument and grade distribution is obtained. Likewise, Botelho et al. (2015) compare teacher assessments of 8th graders across 10.6 thousand classrooms in Brazil to standardized scores (blindly marked) to study racial discrimination. Using the length of classroom interaction time between the teacher and a student as a proxy for individual-level information, they show no racial discrimination for students graded by a teacher who had already taught them, with discrimination only being present for those attending classes with a new teacher. Studies on the impact of rubrics on assessment also uncover compatible patterns with statistical discrimination: teachers' racial bias in grading is present with vague rubrics but disappears when using a rubric with 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).

A key implication differentiating the statistical discrimination theory from those discussed above is the expectation that the more (less) information provided, the less (more) discrimination exists. Applied to the educational context, we argue that statistical discrimination is unlikely in short-term outcomes, like specific grading tasks and retention recommendations, where teachers count on concurrent comprehensive information for accurate assessments (Wenz and Hoenig 2020). By contrast, we predict statistical discrimination to be in place when teachers express individual long-term educational expectations, as they lack crucial information about time-varying factors (e.g., performance) conditioning students' future trajectories. In such an ambiguous case, teachers might rely on time-constant student-ascribed characteristics as a proxy to make informed predictions (Geven et al. 2021). Teachers do not necessarily need in-service experience to infer the actual average success probability of different ascribed groups because this common-ground knowledge might already be acquired during their schooling and pre-service training.

In sum, the expectation we derive from statistical discrimination theory is that ascribed characteristics are more likely to be relevant in evaluations when the performance information is more unreliable (Aigner and Cain 1977) or conveys less diagnostic clarity (Fiske et al. 2018). Thus, we expect that teachers are more likely to rely on ascribed characteristics when the information is unreliable and uncertain—which is the case in long-term assessments—than when the information is reliable and certain—the case in short-term evaluations (Wenz and Hoenig 2020). Then, when information is ambiguous, group ascriptive factors (stereotypes) should gain weight as proxies of mean potential success *vis-à-vis* individual ability factors signaling current performance, eventually leading to higher (statistical) discrimination in future than short-term outcomes.

Hypothesis 2 (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). Discrimination is generally larger and consistent with statistical discrimination 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.

Cultural Capital

Cultural capital theories (CCT) define cultural capital as high-status cultural signals that enhance social inequality (Bourdieu and Passeron 1990; Bourdieu 1984). Despite its early influence, cultural and stratification sociologists agree on its unclear formalization and causal basis (Jæger and Breen 2016; van de Werfhorst 2010; Goldthorpe 2007; Lamont and Lareau 1988). Despite its shortcomings, cultural capital can be a powerful educational inequality mechanism if precisely formalized and tested (Jæger 2022). Thus, this article seeks to test whether CCT can explain teacher discrimination in addition to the theories reviewed above.

Two different approaches link cultural capital to social stratification. The first argues that cultural capital shapes educational inequality via teachers' bias (Bourdieu 1984). Teachers positively evaluate those children socialized in the dominant culture to which teachers belong, and the school system legitimizes via canonical curricula (Bourdieu and Passeron 1990). Through embedded cultural scripts as 'frames' or 'narratives' (Lamont et al. 2014; Lamont and Small 2008), teachers interpret cultural capital as signals of student academic brilliance *independently* of their actual ability (Jæger et al. 2023). Thus, teachers use and recognize such signals to gatekeep school progress, resulting in cultural discrimination (Jæger and Breen 2016; DiMaggio 1982).

The second approach departs from the classic Bourdieusian proposition to conceive cultural capital as a set of socio-emotional or non-cognitive skills (Farkas 2003)—"patterns of thought, feeling and behavior" (Borghans et al. 2008:974). In turn, these skills directly improve educational performance (Breinholt and Jæger 2019) or the capacity to command attention and negotiate advantages in the classroom (Calarco 2014; Lareau 2011). Thus, those children (and parents) who display high cultural capital also tend to have high ability and motivation, potentially overestimating the effect of cultural capital (Jæger 2022).

Distinguishing between these two approaches to cultural capital is empirically challenging, questioning the causal relationship between cultural capital and educational outcomes (Jæger 2022). This article aims to disentangle these two perspectives linking cultural capital to student academic 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 cultural capital signals, not ascribed characteristics per se, drive teacher discrimination. Even if high-SES or non-migrant background students are more likely to show such signals in the real world, a portrayal of cultural capital signals drives teachers' discrimination.

Hypothesis 3 (H3). Teachers misconceive academic brilliance with highbrow cultural capital by over-grading (H3a), recommending less grade retention (H3b) and expressing higher expectations (H3c) for students signaling high cultural capital (vs low cultural capital).

Data, Variables, and Methods

Sampling Design and Data

Experimental studies generally collect convenience samples that are not representative of the reference population and have low power and external validity. To address these issues, we implemented an explicitly stratified systematic random sampling by public and private institutions with probability proportional to size (see online supplement Part B for details). We randomly drew 20 institutions—15 public and five private—from the population frame to represent all education faculties across non-bilingual Spanish regions ($N = 85$).[3] We replaced four out of the five 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 four private institutions participated in the study, inviting 27,015 students enrolled in the BA Degree in Primary Education in 2022/2023 (see online supplement B and Table S.2.) via faculty's email (see online supplement A). 1,028 students in 15 public and 720 in 4 private faculties completed the online survey between April and June 2023. We collected 1,748 observations (7 percent response rate), reduced to 1,717 after excluding fraudulent or underage (age < 18) cases (Gil-Hernández et al. 2023).[4] The pre-registered power analysis shows the analytical sample and most estimated coefficients lie above powered thresholds (see online supplement Part C) (Freitag and Schuessler 2020; Dziak, Collins, and Wagner 2013).

As seen in Table 1, the share of students in public and private institutions is virtually the same in the experimental sample (40 percent) relative to administrative data on the whole reference population (Ministerio de Universidades 2023).[5] Sociodemographic characteristics of our experimental sample are generally balanced when compared to the population, even though there is a slight overrepresentation of females (+9.9 percent), foreign-born (+3.4 percent) and older students

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 enrollment 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.

(+9.8 percent), and an underrepresentation of highly-educated backgrounds (−10.7 percent). As shown in the online supplement (Part I), we successfully replicate the main models adjusted for calibration weights using raking estimators to adjust the population shares of the main individual-level socio-demographic variables (see Table S.7.).

Table 2: Factors, levels and signaling.

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 origin	1. Spanish origin (native majority) 0. Moroccan origin (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

Methods and Variables

Experimental Design

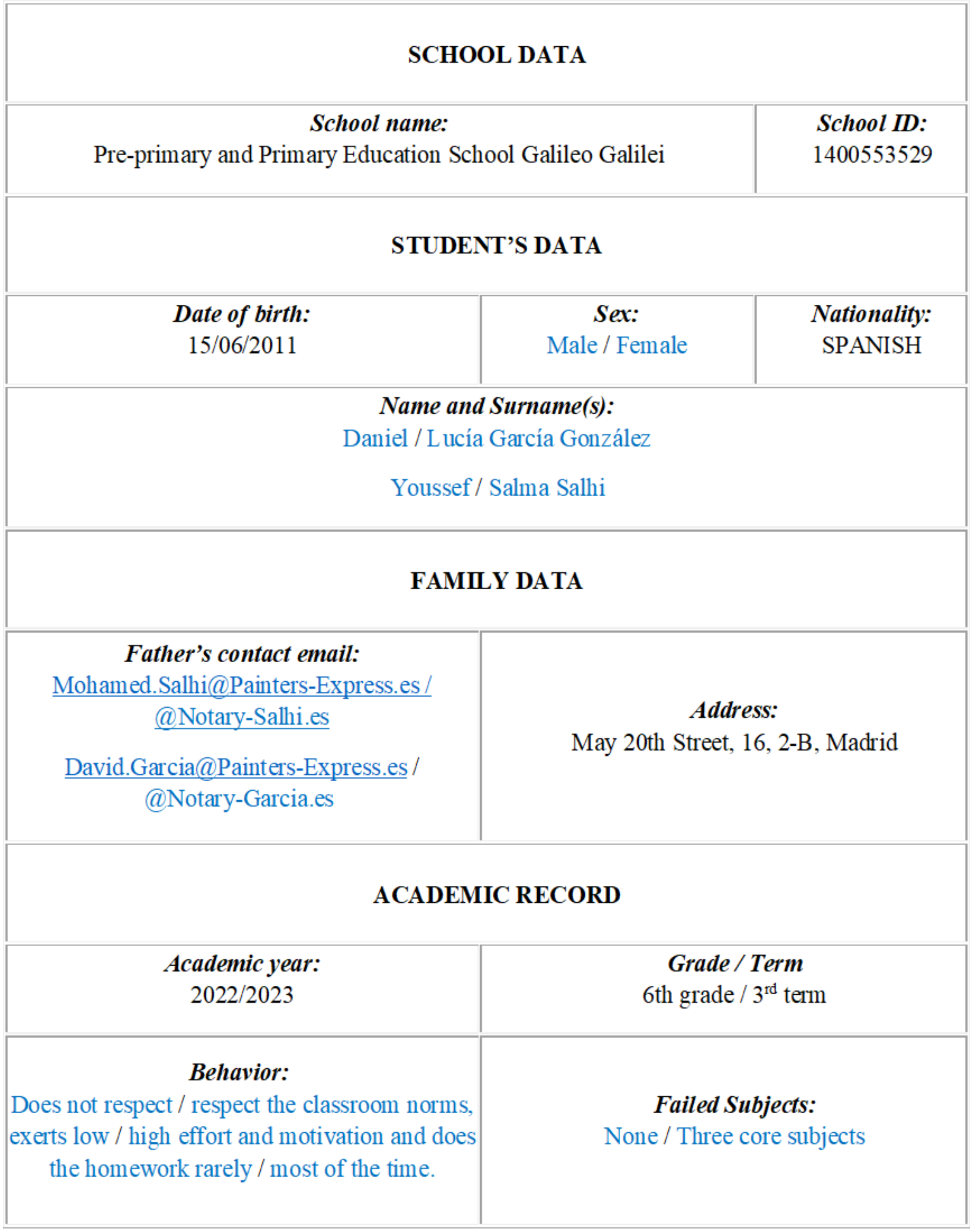
We designed a factorial experiment with $2^7 = 128$ profiles or vignettes—7 dimensions and 2 levels (see online supplement Part A for details on the general set up). As shown in Table 2, we experimentally manipulate student-ascribed characteristics—gender, migrant origin, parental SES, and cultural capital. Besides, to avoid omitted variable bias, we add three dimensions accounting for students' ability and behavior: student language-related skills, subjects failed, and socio-emotional skills (Ferman and Fontes 2022). In section ., we detail how we signal and operationalize each factor. The vignette universe consists of 128 profiles that are orthogonal by design. We implement a full factorial design, including all possible combinations to minimize standard errors and maximize estimation precision. This allows for identifying all main effects independently of each other and all two-way interactive terms, exploiting the maximum variance (Auspurg and Hinz 2015). Potentially non-realistic or implausible combinations in empirical terms are not excluded to avoid loss of efficiency. Only one vignette or task by each respondent was assigned—a between-subject design, to avoid cognitive overload given response times identified in the pre-tests, learning heuristics, and measurement error. The vignettes are the analysis unit and randomly assigned to respondents. To avoid confounding the experimental conditions with respondent characteristics, each vignette was rated by 14 different respondents, on average. Respondent-level covariates are included to increase the precision of the estimates (see section). We ran collinearity tests among factors indicating successful randomization (see online supplement Part G). The online experiment and experimental conditions randomization were implemented with *Qualtrics*® software. Online supplement B details the structure and screens composing the online questionnaire.

Factorial Manipulations: Measurement and Signaling Instruments

This study avoids social desirability bias by hiding the true scope of the research while using realistic signaling instruments: a table resembling a student's file and a digitally transcribed essay written by real students. First, as shown in Figure 1, we built a fake but realistic student file, replicating those used by Spanish in-service teachers to signal student-ascribed factors to assess discrimination and objective ability indicators.[6] To hide the experiment aim, we added five factors usually included in student files, keeping them constant across respondents (e.g., school, academic year, age, nationality and family address). Second, we used an essay varying by its objective externally validated quality to signal students' objective language ability and cultural capital. We subtly embedded cultural capital signals and reinforced 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. We ran manipulation checks to ensure participants were exposed to the treatments (see online supplement part F, Figure S.4.), with more than 80 percent correctly recalling every single factor and 60 percent all factorial manipulations. We replicate the main analyses on a subsample of respondents correctly recalling all treatments (see online supplement Part I, Table S.6.).[7]

Gender. Student names vary by gender and migration background. We select the most common and region-neutral (no names from bilingual 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 2023b). For Spanish origin students, boys are named *Daniel* (0) and girls *Lucía* (1). For Moroccan origin (see below) pupils, the boys' name is *Youssef* (0), and the girls' is *Salma* (1). Gender is signaled in the student file and the essay's screen instructions.

Migrant origin. We signal migrant origin with the student's and father's first and last names. We picked the most common Spanish and foreign origin surnames among newborns in 2011 for children and fathers (INE 2023a). For Spanish origin: (1) *García* and *González*; for foreign origin (Moroccan): (0) *Salhi*. Among fathers, we chose the most common father's name according to the INE for those born in the 1980s. For Spanish origin (1): *David*; for foreign origin (Moroccan): (0) *Mohamed*. Migration origin is signaled in the student file (student and father) and



SCHOOL DATA

School name: Pre-primary and Primary Education School Galileo Galilei	***School ID:*** 1400553529

STUDENT'S DATA

Date 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

Figure 1: Student's file example: Experimentally manipulated factors and levels (in blue) and fixed information (in black) in the vignettes (translated from Spanish). *Notes:* Fixed information in black: (1) academic year (2022/2023), evaluation term (last), and grade (sixth); (2) the birth date (15/06/2011) and age (11-12) signaling no previous retention; (3) fake school name and administrative ID with a neutral school type; (4) fake family address with neutral information about the type of house and area; (5) and student Spanish nationality to signal that all students from ethnic minority origin are second-generation.

the essay's vignette instructions (student). Because we could only include two levels for this factor for statistical efficiency, we chose Moroccans as the reference ethnic minority. First, Moroccans represent the largest foreign origin minority, with 28.9 percent (19.6 percent) of primary (lower-secondary) education students (Ministerio de Universidades 2023). Second, among the largest ethnic minority groups in the Spanish school system, Moroccans are the most socioeconomically and academically disadvantaged (Gil-Hernández and Gracia 2018) with the most negative stereotypes (Martínez de Lafuente 2021). Third, Moroccan origin names and surnames are a more powerful signal in Spain than Romanian or Latin American.

Parental SES. We signal students' SES through the father's occupation in the student's file and embed it in the essay (see below). We selected parental occupations (fathers to control for family structure) commonly perceived as high or low-SES or prestige by 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 subtly signaled parental SES in the student's file family contact module (see Figure 1) through the father's email (Martínez de Lafuente 2021), including name, surname, and occupation/business. For the high-SES occupation (notary), we included the father's surname in the email domain to elicit that he owns a small-to-medium notary firm to prevent SES underestimation among ethnic minority fathers compared to the native majority (Crabtree et al. 2022). For the low-SES occupation (painter), the father's surname was not included in the email domain to elicit he is an employee within the firm. 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, orthogonal to its quality and cultural capital, that 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 Tables S.3.-S.4. for details).

Cultural capital. We signal embodied cultural capital (Sullivan 2002) within the student's essay (see online supplement Tables S.3.-S.4.). 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 low-brow leisure activities that convey different social statuses, recognition, or legitimacy in the dominant cultural hierarchy (Jæger and Larsen 2024; Jæger, Rasmussen, and Holm 2023; Bourdieu 1984). Low cultural capital (0) is signaled through a low-brow or popular leisure activity referenced in the essay: watching a reality show on television (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 (Jæger et al. 2023). To ensure respondents perceive the embodied cultural capital signals, we successfully pre-tested its internal validity with 243 in-service elementary education teachers (see online supplement Part E for details).

Language ability: objective essay quality. We randomly assigned two versions of a short essay varying in its objective quality (0=bad; 1=good) regarding structure,

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

orthography, vocabulary, and creativity to capture students' language ability (see online supplement Table S.3.). We asked real sixth graders to write essays about a neutral topic (a landscape of their liking) regarding region and ascribed characteristics, which were digitally transcribed. As an external objective quality benchmark (Quinn 2020), we applied official Spanish competence rubrics for the elementary 6th grade. We pre-tested the objective grade assigned to the digital essay using a sample of 243 in-service elementary education teachers (see online online supplement Part D). As Table 3 displays, our experimental sample, consisting of pre-service teachers, 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 (pooled mean of 7.3 and SD=2), indicating high internal validity. Thus, compared with the in-service teachers' pre-test, our measure of student language ability shows high external validity.

As shown in the online supplement Part D and Table S.3., 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 further 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); all subjects passed (1). Student academic performance is signaled in the student file.

Socio-emotional skills. To capture students' 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 2022; 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.

Table 4: Outcomes' summary statistics (upper panel) and correlation matrix (bottom panel)

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

Correlation

	(*a*)	(b)	(*c*)
Essay Grade (a)	1		
Grade Retention (b)	−0.54	1	
Academic Track (c)	0.42	-0.42	1

Outcomes

All outcomes are measured in a metric scale to maximize variation. Table 4 below reports the summary statistics of the outcomes. In online supplement Part I, we run a robustness check to account for the non-normal distribution of the outcomes (see Table S.8.).

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. Still, its prevalence in 2020 was 2.3 percent, considerably above the OECD average of 1.3 percent.

Educational expectations. about enrollment 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 it is not at all likely to happen, and 10 is very likely.* 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.

Estimation and models

In the baseline set of models (M1) (only shown in online supplement Table S.5B), we run Ordinal Least Squares (OLS) models, 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, the causal estimand of interest, expresses a factor's average individual-level effects (Hainmueller, Hopkins, and Yamamoto 2014). It can be interpreted as the causal effect of a specific factor level (treatment) in comparison with another level of this same factor (baseline or control category) while keeping equal the joint distribution of the remaining factors (Bansak et al. 2021). The remaining factors operate as randomized pre-treatment covariates. Standard errors are clustered at the faculty/university level to account for the non-independence of observations within these sampling groups. In the second set of models (M2), formalized in Equation 1 and fully displayed in online supplement Tables S.5A-S.5B, pre-treatment respondent-level controls (ethnic origin, gender, parental SES, grade retention, institution fixed effects, BA enrollment grade and birth year) are a covariate vector ($\mathbf{Z}_j$) to increase the precision of the main effects because individual-level variables are independent of the experimental factors by design (Baguley, Dunham, and Steer 2022).

$$\begin{aligned} y_{i123} = \alpha &+ \beta_{i1}\ gender + \beta_{i2}\ migrant\ background + \beta_{i3}\ parental\ SES \\ &+ \beta_{i4}\ cultural\ capital\ + \beta_{i5}\ essay\ quality + \beta_{i6}\ subjects\ failed \\ &+ \beta_{i7}\ socioemotional\ skills + \mathbf{Z}_j + \varepsilon_i \end{aligned} \tag{1}$$

According to the pre-registered power analysis and final analytical sample we reached in the study, we do not test the moderation hypotheses outlined in the PAP. We only run interaction models in two specific cases relevant to the main hypotheses outlined above, where large and powered heterogeneous effects can be identified by essay quality ($y_{i1essay\ grade}$) and failed subjects ($y_{i2retention}$). We provide two additional non-pre-registered analyses in the PAP to better disentangle our research hypotheses. First, at the bottom of online supplement Table S.5A, we include the ratio between ability and ascriptive factors, dividing the average absolute effect size of the three ability factors by the average absolute effect size of the four ascriptive factors by each outcome. An ability/ascriptive ratio considerably smaller for long-term expectations than the remaining short-term outcomes would align with *H2* on statistical discrimination. Second, instead of experimentally assigning information treatments, we hold individual student information constant within respondents to exploit variation in its reliability and uncertainty across short- and long-term outcomes. The same student information assigned to each teacher should convey less reliability the more the outcome is projected into the future. In Figure 3 (full output in online supplement Table S.11.), we formally test differences in the discrimination coefficients across outcomes with a two-tailed Z-test from seemingly unrelated regressions that account for covariance between estimators within the same sample (Clogg, Petkova, and Haritou 1995). In this way, we assess whether student's ascribed characteristics more strongly impact long-term educational expectations ($y_{i3expectations}$) when compared to the models predicting short-term outcomes ($y_{i1essay\ grade;}\ y_{i2retention}$). If so, *H2* would gain additional support.

Findings

Figure 2 portrays the main OLS models (M2) output by outcome and experimental factor, controlling for sampling institution fixed effects and respondent characteristics (full output in online supplement Tables S.5A-S.5B). In Figure 2, in the upper panel, we split the independent variables by *ascribed* and *ability* factors. These factors represent multiple randomized categorical treatments concerning its control or reference group. Figure 2 (upper panel) shows that those factors accounting for students' objective ability are the most predictive *vis-à-vis* ascriptive factors across outcomes. Unsurprisingly, students' cognitive and non-cognitive skills are the leading performance indicators for teachers' assessments.

Essay Grading

Focusing on grading, an objectively *good* essay implies 2.8 points (p value <0.001) higher teacher grading than an objectively *bad* essay. The remaining ability factors, number of failed subjects and socio-emotional skills, have similar predictive power, with AMCEs at 0.28 (p value <0.01) and 0.27 (p value <0.01), respectively. In grading an essay, a student's number of failed subjects or behavior is arguably outside the scope of what should be graded according to official Spanish rubrics for language tasks. This finding mirrors previous research showing that teachers assess students according to their classroom behavior beyond their objective competence (Ferman and Fontes 2022) and the importance of controlling for it to identify teacher bias.

Figure 2 (bottom panel) zooms in on the particular role of ascriptive factors. Net of students' objective observed ability, teachers tend to assign higher grades in the essay, on average, to students profiles who are girls ($\beta_{\text{AMCE}} = 0.12$ [p value <0.1]), come from a (Moroccan) ethnic minority origin ($\beta_{\text{AMCE}} = 0.2$ [p value <0.01]), or signal high cultural capital ($\beta_{\text{AMCE}} = 0.2$ [p value <0.001]). This latter finding on cultural capital discrimination validates *H3a*: direct exposure to a written highbrow cultural capital signal in a real student's essay task elicits higher teacher evaluations. In line with *H3a*, 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. The effect is nevertheless limited as it does not show on the remaining outcomes, rejecting *H3b-c*. Moving on to student parental SES, this factor is irrelevant in predicting teacher grades, partially rejecting *H1a* on implicit bias or SCT. In the discussion, we elaborate on this null finding compared to previous experimental and observational studies. For the case of student gender, the magnitude and direction of the coefficient point to slight positive grade discrimination for girls, as expected by *H1a* and broadly in line with previous findings. Yet the estimation is highly uncertain with a p value < 0.1. Online supplement Figure S.7. reveals that girls are particularly over-graded in comparison with boys when the essay is good ($\beta_{\text{AMCE}} = 0.27$; p value < 0.05; $n = 871$), backing

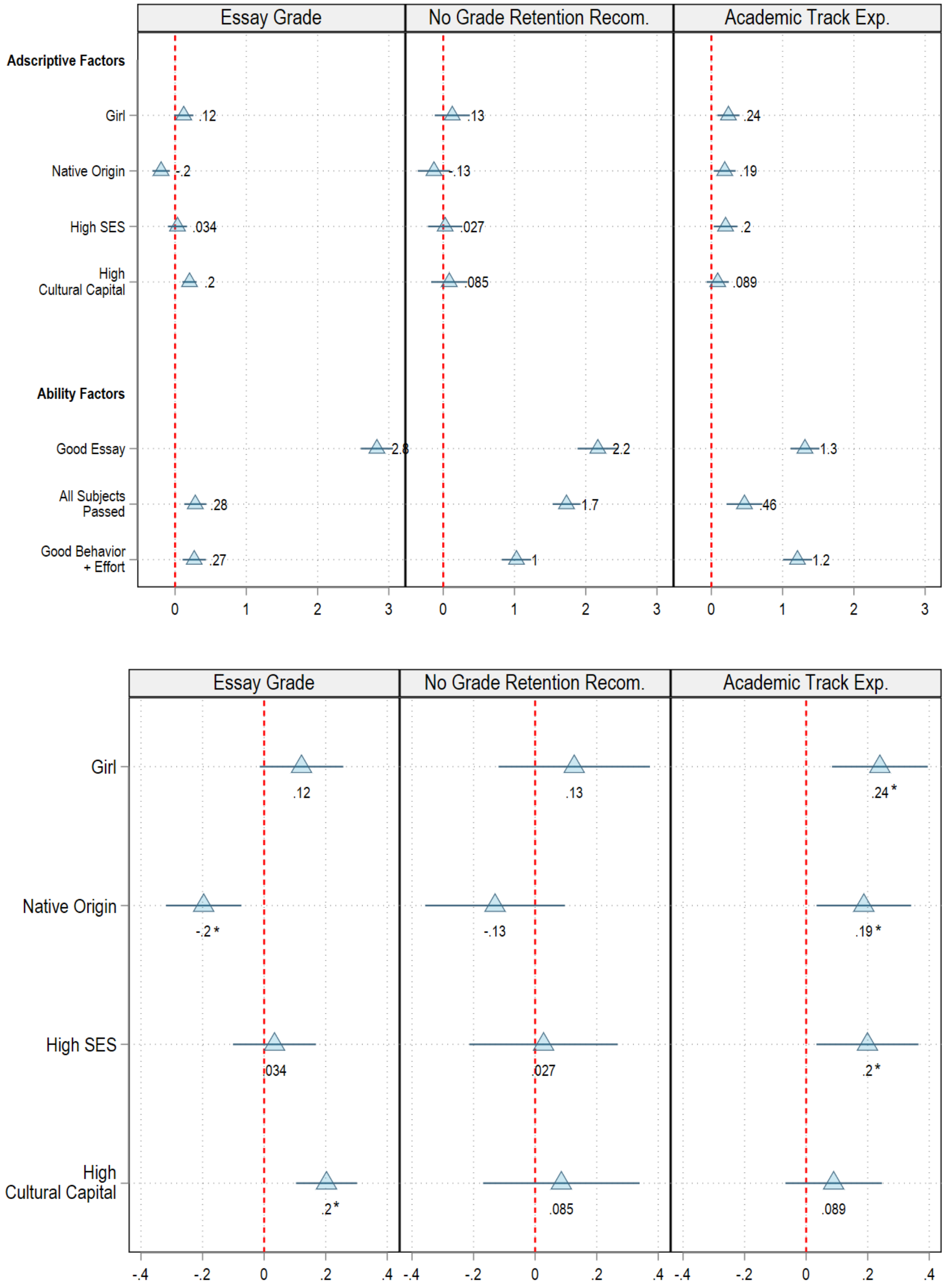


Figure 2: AMCEs of ascriptive and ability factors on educational outcomes (95% CI). *Notes:* *p value < 0.05 (Bottom panel); Controls: institution-FE; respondents' characteristics. The bottom panel displays the same model as the upper, zooming in on ascriptive factors. The full output is in online supplement Tables S.5A-S.5B. The grade retention scale is reversed in this figure to display the same coefficients' direction across outcomes and facilitate interpretation. Recom. = Recommendation; Exp. = Expectations.

the generalized belief that girls are more competent than boys in language tasks (Homuth et al. 2023). In turn, the finding on positive ethnic discrimination in essay grading, contrary to *H1a* and most previous findings, is surprising. We did not expect this coefficient's direction favoring ethnic origin pupils under any of the above-discussed discrimination theories. It seems that teachers tend to compensate for a student's overall disadvantaged ethnic minority origin by over-grading them in comparison to the equally skilled and socioeconomically (dis)advantaged ethnic majority. We discuss this unexpected finding in the concluding section 5 in line with compensatory discrimination.

Grade Retention Recommendations

We now turn to the second outcome, teacher recommendations for student grade retention in elementary 6th final grade. As expected, all ability-related factors are statistically significant and highly predictive. 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 ($\beta_{\text{AMCE}} = 2.16$ [p value < 0.001]) of (no) grade retention recommendation. Following in effect size, having passed all subjects ($\beta_{\text{AMCE}} = 1.73$ [p value <0.001]) and students' good behavior and effort ($\beta_{\text{AMCE}} = 1.03$ [p value < 0.001]) in the current term evaluation are considerably more predictive than they were for essay grading. The larger effect size of the student's socio-emotional skills and, notably, the 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*.

Focusing on the student's ascribed characteristics, although the coefficients' direction and effect sizes generally align with our findings for essay grading, none is statistically significant under the standard 5 percent threshold. When interpreting these *a priori* null findings, one should consider the outcome's high variation and skewness to the right (see online supplement Figure S.8.), seemingly indicating that most teachers are highly averse to grade retention. Thus, we run a heterogeneous model (M2) by the number of failed subjects to test for a more realistic setting. As illustrated by online supplement Figure S.9., we find strong positive gender ($\beta_{\text{AMCE}} = 0.36$ [p value < 0.05]) and ethnic minority ($\beta_{\text{AMCE}} = 0.53$ [p value < 0.01])—compensatory—discrimination in (no) grade retention recommendations, in line with and contrary to *H1b*, respectively. In turn, for parental SES and cultural capital, *H1b* does not find support.

Upper-Secondary Track Expectations

The third outcome is teacher expectations about a student's enrollment in the upper-secondary academic track. In this long-term outcome, teachers evaluate students' profiles by considering the last grade of elementary education, lacking information about future performance. After that, lower-secondary education comprises four grades before the end of compulsory schooling (16 years old) and the transition into upper-secondary education.

As evident from its smallest adjusted coefficient of determination (R^2) across outcomes (see online supplement Table S.5A), educational expectations seem more uncertain for respondents, signaling a noisier model. On average, ability factors have smaller effect sizes for long-term expectations, with the smallest average ratio over ascriptive factors (5.6), halving the remaining outcomes (12.7). Still, student behavior and effort ($\beta_{\text{AMCE}} = 1.21$ [p value < 0.001]) seem to gain weight compared to the previous outcomes as a powerful indicator of future success. Hence, students' current performance is not fully informative for teachers to predict future attainment accurately, leaving room for ascribed group-level stereotypes.

Focusing on ascribed factors in Figure 2 (bottom panel) reveals that, in line with *H2* and previous findings (Geven et al. 2021; Timmermans et al. 2015), teachers express higher long-term expectations for those groups with historically higher educational attainment across secondary education (Gil-Hernández and Gracia 2018), such as girls ($\beta_{\text{AMCE}} = 0.24$ [p value <0.01]), native origin ($\beta_{\text{AMCE}} = 0.19$ [p value <0.05]) and high-SES ($\beta_{\text{AMCE}} = 0.2$ [p value <0.05]) students, displaying similar effect sizes at about 10 percent of an SD-unit.

As predicted by *H2* and formally tested in Figure 3 with coefficient difference tests by outcomes (see online supplement Table S.11.), the effect sizes of ascriptive factors like gender ($\Delta\beta_{\text{Expectations-Grading}} = 0.05$; $\Delta\beta_{\text{Expectations-Retention}} = 0.07$), parental SES ($\Delta\beta_{\text{Expectations-Grading}} = 0.08$; $\Delta\beta_{\text{Expectations-Retention}} = 0.08$), and ethnic origin ($\Delta\beta_{\text{Expectations-Grading}} = 0.19$; $\Delta\beta_{\text{Expectations-Retention}} = 0.13$) are substantially larger, from 5 percent to 20 percent an SD-unit—even changing sign for ethnic origin, for long-term educational expectations than for short-term outcomes. Nevertheless, the z-tests yield statistically significant differences (p value < 0.05) only for ethnic backgrounds. Although potentially underpowered, these findings align with statistical discrimination as student-ascribed characteristics gain weight in teacher long-term evaluations, increasing discrimination as information on student ability becomes less reliable.

At the same time, these findings on bias in educational expectations against boys, low-SES and ethnic minority students are compatible with *H1c* on implicit bias and status characteristics beliefs. These theories predict teacher bias to remain relatively independent of the level of individual information dispensed, as teachers will stick to pre-existing status beliefs to form their competence expectations by ascribed groups.

Discussion and Conclusions

Fair evaluations are critical for equal educational opportunity. Teachers are the principal evaluators of academic merit in the educational system. Nevertheless, their direct role in reproducing or compensating educational inequalities remains elusive as previous observational work and the few experimental studies available have yielded inconclusive. Thus, this article tested if (pre-service) teachers show discrimination in their assessments and expectations as a function of student-ascribed characteristics with a causal experimental design.

We framed our research hypotheses from multidisciplinary theories of status characteristics beliefs, implicit bias, statistical discrimination, and cultural capital.

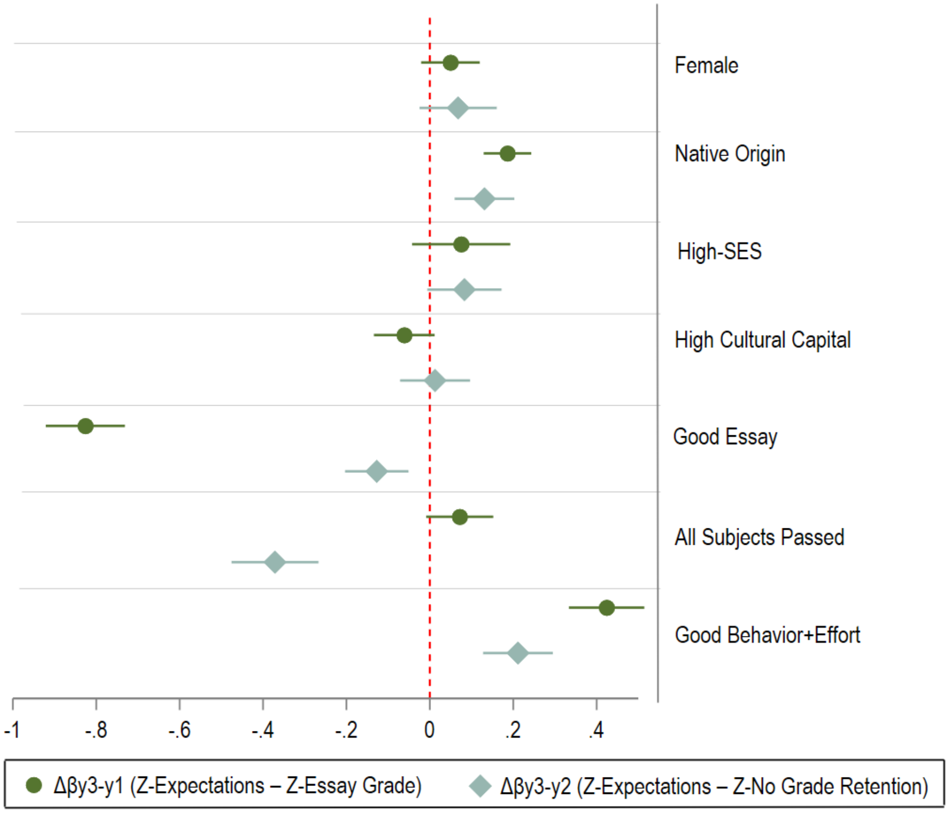


Figure 3: Standardized coefficient differences across outcome models (95% CI). *Notes:* Coefficients difference by outcome models ($\Delta\beta y3 - y1$ [Z-Academic Track Expectations – Z-Essay Grade]; $\Delta\beta y3 - y2$ [Z-Academic Track Expectations – Z-No Grade Retention]) with a two-tailed Z-test using seemingly unrelated regressions. Reversed scale for grade retention recommendation outcome to accurately estimate coefficient differences. Outcomes in z-scores for scale comparability. Clustered standard errors. Full output in online supplement Table S.11.

We analyzed different outcomes over the students' educational careers, conveying diverse uncertainty for teacher evaluations to disentangle these theories' predictive power. We conducted a pre-registered full factorial survey experiment with realistic and externally validated instruments, drawing a large sample of Spanish pre-service teachers before exposure to the school context. For the first time, this research design causally identifies the net effect of different student-ascribed characteristics—gender, SES, ethnic origin, and cultural capital—beyond ability on teachers' assessments.

Online supplement Table S.12. summarizes the article's main findings. Overall, we found teacher biases in (essay) grading favoring girls (supporting *H1a* on *status characteristics beliefs and implicit bias theories*), ethnic minority origin (partially rejecting *H1a*), and students signaling high cultural capital (partially supporting *H3a* on *cultural capital 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-performing students falling within the legal threshold for repeating a grade. Finally, concerning teachers' educational expectations, we found evidence of *statistical discrimination* (validating *H2*) in favor of girls, native origin students, and high-SES background students.

For essay grading, a short-term outcome where we allegedly provided teachers with the minimum necessary information for fair assessments, results align with theories of *implicit bias* or *status characteristics beliefs* (gender), *cultural capital* (signals), and previous findings. The finding on gender supports the generalized belief that girls are more competent at school (Homuth et al. 2023), as they objectively overperform boys, especially in language competencies like the ones evaluated in this experiment essay task.

The finding on the effect of cultural capital on essay grading, not holding for grade retention recommendations or long-term expectations, suggests that immediate exposure to a highbrow culture signal boosts teacher perceptions of academic brilliance beyond the student's true ability (Jæger and Møllegaard 2017). Cultural capital is orthogonal to student SES by design in this experiment. Still, given that they positively associate in reality, the former might be a causal mechanism driving SES-based inequality in assessments.

Contrary to our expectations in *H1a-b*, we find no evidence of SES bias 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 et al. 2022) for not fully controlling for socio-emotional skills (Ferman and Fontes 2022) and/or measurement error in test scores (van Huizen et al. 2024); (2) underestimation in our essay grading task due to low ecological validity because, 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, contrary to most previous studies (Zanga and De Gioannis 2023; Gortázar et al. 2022) and *H1a-b*, we found evidence of *over-grading* and *under-expectations* of grade retention for ethnic minority students, suggesting explicit *compensatory discrimination*. 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 teachers' equalizing preferences. In our investigation, absolute grading practices should prevail (Hjorth-Trolle et al. 2022) because each respondent only evaluated one student profile. Still, despite student performance and SES being 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. Furthermore, about 80 percent of second-generation Moroccan origin students do not regularly speak Spanish at home, being one of the most socioeconomically deprived minorities (Gil-Hernández and Gracia 2018:594). Thus, teachers might generally perceive that they are a disadvantaged minority experiencing language difficulties and, hence, 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, whereas they do affect math, arguing that literature teachers might internalize the need to help immigrants less acquainted with the Italian language, regardless of their biases (Alesina et al. 2018:3). Supporting this pattern, in Spain, Marcenaro-Gutiérrez and Vignoles (2015) found higher teachers' grades for migrant origin students than their performance in a blind standardized reading test, relative to the Spanish origin majority, but not in math.

The observed biases in educational expectations of upper-secondary pathways—a long-term outcome lacking information on students' future performance—favoring girls, native origin students, and high-SES students support *statistical discrimination theories*, validating *H2*, and in line with previous experimental findings on in-service teachers (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. Simultaneously, these findings on educational expectations also validate *H1c* on implicit bias and status characteristics beliefs, predicting teacher bias to be cognitive in origin and relatively independent of the degree of individual information disclosed (Correll and Benard 2006).

The finding on (negative) statistical 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 enrollment choices (Ferrara 2023), compared to equally-performing peers from national majority origin. Thus, beyond being educated guesses, statistical discrimination practices might lead to self-fulfilling prophecies 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).[8]

On average, as the benchmarking analysis in online supplement Part L shows (see Table S.12.), we reported average effect sizes (Cohen's D $\approx$ 0.1) in line with previous studies (Schuessler and Sønderskov 2023; Alesina et al. 2018) representing more than 50 percent of learning gains over a school year (Evans and Yuan 2019), large-scale educational interventions, or gender gaps in test scores. These benchmarks indicate teacher bias effects are not trivial and might entail real consequences for educational pathways, especially when accumulating (dis)advantages over several assessments (DiPrete and Eirich 2006). 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.

Limitations and Conclusion

This study has four limitations that pave the way for future research. First, a complex trade-off exists between avoiding social desirability and ensuring respondents internalize the experimental manipulations in ecologically valid factorial designs.

We implemented externally validated survey instruments that were as realistic as possible to emulate real world evaluation settings. However, fictitious student profiles might trigger statistical discrimination, as in-service teachers know their students, which reduces the lack of individual-level information. Still, as shown by Krolak-Schwerdt et al. (2017), vignettes of fictitious students yield ecologically valid results of teachers' assessments in real classrooms. 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, whereas our vignette experiment induced absolute grading in a single task. Absolute and relative grading scales might have different implications for students' ascribed status inequalities depending on school composition (Hjorth-Trolle et al. 2022), whereas teachers' biases might accumulate over several evaluations to assign the final grade. Finally, given implicit biases against girls' scientific competence (Carlana 2019), implementing a math task might go more in line with classic SCT predictions of men being expected to outperform women (Correll and Ridgeway 2006) and not replicate our findings on teacher's over-grading of girls in essays. Field school experiments combining administrative data on fully comparable internal and external grades (Bygren 2020) in language and scientific subjects and automated cognition tests might overcome these challenges (Alesina et al. 2018).

Second, our sample of pre-service teachers raises external validity issues, as most have not yet had direct contact with students or schools. Pre-service teachers are a more homogenous group than in-service teachers in terms of age and experience, and, as younger cohorts, they might be more idealistic and unaffected by real-school practice. Then, one might wonder to what extent the findings reported here represent a lower- or upper-bound of what we would find with actual teachers. Because pre-service teachers lack real experience, their evaluation practices might not be accurate. Still, in a pre-test to validate the instrument, we have shown that, in grading an essay, the grade distribution between pre- and in-service teachers largely overlaps. For retention recommendations, one could expect pre-service teachers to be more idealistic and averse than older in-service teachers, especially when this practice is currently discouraged by national and international educational institutions. Regarding long-term educational expectations, fictitious student profiles might boost statistical discrimination, especially among pre-service teachers who lack experience teaching individual students, as they might *fill in the blanks* by assigning the corresponding ascribed group stereotype. Yet the factorial design with fictitious student profiles might mitigate this experience gap between pre-and in-service teachers, being similarly affected by variation in student information uncertainty.

Inter-group relations theories predict less (*contact theory*) or more (*conflict theory*) discrimination as a function of *inter-group* exposure, yielding mixed and scarce findings in the educational context (Elwert, Keller, and Kotsadam 2023). Thus, one can expect pre-service teachers to be either less or more biased by student-ascribed factors than in-service teachers with field experience. Prior 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). Our study also identified comparable effects to observational and experimental studies with in-service teachers (Schuessler and Sønderskov 2023; Geven et al. 2021; Wenz and Hoenig 2020; Alesina et al. 2018; Marcenaro-Gutiérrez and Vignoles 2015). Contrary to *contact* and *conflict theories*, a field experiment on ethnic discrimination among Hungarian students (Elwert et al. 2023) indicated that randomly manipulating inter-ethnic exposure or ethnic composition within classrooms did not affect peer 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. Starck et al. (2020) have demonstrated 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. Future studies can test whether our findings generalize to other national contexts or replicate with in-service teachers, testing *inter-group relations* theories. Our motivation to focus on pre-service teachers was that in-service teachers might sort into schools with practices and student compositions aligned with their previous biases, which school-level factors might reinforce or mitigate (Pit-ten Cate and Glock 2019). Thus, an open empirical question is whether estimations on pre-service teachers are externally valid or can establish a benchmark for inter-group relations studies.

Third, we applied a random sampling design to cover our frame population, reaching a larger and more representative sample than most previous experimental studies on educational discrimination run on convenience samples. Furthermore, we pre-registered a power plan to identify powered effects and bypass most previous underpowered studies. 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, given the benchmark effect sizes and power we reported in this study, we recommend that future studies collect larger samples to more reliably identify potential false negatives and interaction effects.

Fourth, with our factorial design, we cannot causally identify the relative explanatory power of different theories and mechanisms, as we did not randomly assign various degrees of student information to teachers or deploy tests of automatic cognition to disentangle implicit bias and SCT from statistical discrimination (Melamed et al. 2019). Nevertheless, we formalized a statistical discrimination test by comparing the relative impact of ascribed and ability student information by educational outcomes with different degrees of uncertainty. Future studies might apply more fine-grained experimental designs to better untangle these mechanisms. That is not an easy task because implicit bias, SCT, cultural capital, and statistical discrimination are not competing theories and might operate simultaneously (Correll and Benard 2006).

Having acknowledged these limitations, we showed for the first time the causal effect of several ascribed characteristics—gender, SES background, ethnic origin, and cultural capital—among equally competent students on (pre-service) teacher assessments. We uncovered complex bias dynamics, expanding our knowledge of discrimination as a relevant educational inequality mechanism. Consciously or not, teachers perceived some ascribed groups of students as more competent, deserving, or likely to succeed than others, despite equal objective performance. That leads to biased assessments in a fictitious experimental setting that might translate into self-fulfilling prophecies and cumulative (dis)advantages over actual schools. 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 implicit biases against immigrants might not align with explicit judgment behavior. We hope 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.

Notes

[1] For simplicity, from now on, we refer to cultural capital as an ascribed characteristic in addition to gender, migrant origin, and socioeconomic background. In section , we formalize the concept of cultural capital.

[2] In the pre-registered pre-analysis plan, we did not specify different hypotheses for the empirical expectations expressed here as hypotheses 1, 2 and 3 on implicit bias or status characteristics theory (*H1*), statistical discrimination (*H2*), and cultural capital theories (*H3*), respectively. We formalized them jointly in the pre-analysis plan (*H1a-d*), but the exact predictions by students' ascribed factors apply here.

[3] We focus on non-bilingual regions to prevent regional identity and discrimination to confound ascribed characteristics. We also excluded bilingual regions as our task involves Spanish competencies, which might vary by (non)bilingual regions.

[4] We ensured that participants provided honest and accurate responses by running attention checks (to drop those observations who replied too fast or completed the survey randomly) and identifying and filtering out duplicates.

[5] Note that this figure is not directly comparable as, in the administrative data, migrant origin students are defined as non-Spanish nationals. In our experiment, we ask for the parental country of birth.

[6] 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 on two different screens.

[7] Not recalling a treatment might proxy for non-discriminatory behavior as respondents might not consider a given student's ascribed factor relevant information for their assessments.

[8] In online supplement Part J, we test whether participants assume low-SES or migrant origin students will count on less parental support during their prospective education, with null results backing statistical discrimination in long-term expectations.

References

Aigner, Dennis J., and Glen G. Cain. 1977. "Statistical theories of discrimination in labor markets." *Industrial and Labor Relations Review* 30(2):175-187. https://doi.org/10.1177/001979397703000204 https://doi.org/10.2307/2522871

Alesina, Alberto, Michela Carlana, Eliana La Ferrara, and Paolo Pinotti. 2018. "Revealing Stereotypes: Evidence from Immigrants in Schools." *NBER Working Paper* No. 25333. https://doi.org/10.3386/w25333

Arrow, Kenneth J. 1973. "The Theory of Discrimination." Pp. 3–33 in *Discrimination in Labor Markets*, edited by O. Ashenfelter and A. Rees. Princeton, NJ: Princeton University Press.

Arrow, Kenneth J. 1998. "What Has Economics to Say about Racial Discrimination?" *Journal of Economic Perspectives* 12(2):91–100. https://doi.org/10.1257/jep.12.2.91

Auspurg, Katrin, and Thomas Hinz. 2015. *Factorial Survey Experiments*. Thousand Oaks, CA: Sage Publications. https://doi.org/10.4135/9781483398075

Baguley, Thom, Peter Dunham, and Oliver Steer. 2022. "Statistical Modelling of Vignette Data in Psychology." *British Journal of Psychology* 113:1143–1163. https://doi.org/10.1111/bjop.12577

Bansak, Kirk, Jens Hainmueller, Daniel Hopkins, and Teppei Yamamoto. 2021. "Conjoint Survey Experiments." In *Advances in Experimental Political Science*, edited by J. Druckman and D. Green, 19–41. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108777919.004

Batruch, Anatolia, Sara Geven, Emma Kessenich, and Herman G. van de Werfhorst. 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

Bennet, Tony. 2006. "Distinction on the Box: Cultural Capital and the Social Space of Broadcasting." *Cultural Trends* 15(2–3):193–212. https://doi.org/10.1080/09548960600713080

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

Borghans, Lex, Angela Lee Duckworth, James J. Heckman, and Bas ter Weel. 2008. "The Economics and Psychology of Personality Traits." *Journal of Human Resources* 43:972–1059. https://doi.org/10.3386/w13810

Borjas, George J., and Matthew S. Goldberg. 1978. "Biased Screening and Discrimination in the Labor Market." *American Economic Review* 68(5):918–922.

Botelho, Fernanda, Ricardo A. Madeira, and Marcos A. Rangel. 2015. "Racial Discrimination in Grading: Evidence from Brazil." *American Economic Journal: Applied Economics* 7(4):37–52. https://doi.org/10.1257/app.20140352

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

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

Breen, Richard, and John H. Goldthorpe. 1997. "Explaining Educational Differentials: Towards a Formal Rational Action Theory." *Rationality and Society* 9(3):275–305. https://doi.org/10.1177/104346397009003002

Breinholt, Asta, and Mads Meier Jæger. 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, Magnus. 2020. "Biased Grades? Changes in Grading after a Blinding of Examinations Reform." *Assessment and Evaluation in Higher Education* 45(2):292–303. https://doi.org/10.1080/02602938.2019.1638885

Calarco, Jessica M. 2014. "Coached for the Classroom: Parents' Cultural Transmission and Children's Reproduction of Educational Inequalities." *American Sociological Review* 79:1015–1037. https://doi.org/10.1177/0003122414546931

Carlana, Michela. 2019. "Implicit Stereotypes: Evidence from Teachers' Gender Bias." *Quarterly Journal of Economics* 134(3):1163–1224. https://doi.org/10.1093/qje/qjz008

Carlana, Michela, Eliana La Ferrara, and Paolo Pinotti. 2022. "Implicit Stereotypes in Teachers' Track Recommendations." *AEA Papers and Proceedings* 112:409–14. https://doi.org/10.1257/pandp.20221005

Cea D'Ancona, María Ángeles. 2016. "Immigration as a Threat: Explaining the Changing Pattern of Xenophobia in Spain." *Journal of International Migration and Integration* 17:569–591. https://doi.org/10.1007/s12134-015-0415-3

Childress, Clayton, Shyon Baumann, Craig M. Rawlings, and Jean–François Nault. 2021. "Genres, Objects, and the Contemporary Expression of Higher–Status Tastes." *Sociological Science* 8:230–264. https://doi.org/10.15195/v8.a12

Chmielewski, Anna K. 2019. "The Global Increase in the Socioeconomic Achievement Gap, 1964 to 2015." *American Sociological Review* 84(3):517–544. https://doi.org/10.1177/0003122419847165

Clogg, Clifford C., Eva Petkova and Adamantios Haritou. 1995. "Statistical Methods for Comparing Regression Coefficients Between Models." *American Journal of Sociology* 100(5):1261–1293. https://doi.org/10.1086/230638

Correll, Shelley J., and Cecilia L. Ridgeway. 2006. "Expectation States Theory." In *Handbook of Social Psychology*, edited by J. Delamater, 29–51. New York, NY: Kluwer Academic/Plenum Publishers. https://doi.org/10.1007/0-387-36921-X_2

Correll, Shelley J., and Stephen Benard. 2006. "Biased Estimators? Comparing Status and Statistical Theories of Gender Discrimination." In *Advances in Group Processes*, Vol. 23, edited by S.R. Thye and E.J. Lawler, 89–116. Leeds: Emerald Group Publishing Limited. https://doi.org/10.1016/S0882-6145(06)23004-2

Crabtree, Charles, S. Michael Gaddis, John B. Holbein, and Edvard Nergård Larsen. 2022. "Racially Distinctive Names Signal Both Race/Ethnicity and Social Class." *Sociological Science* 12:454–472. https://doi.org/10.15195/v9.a18

DiMaggio, Paul. 1982. "Cultural Capital and School Success: The Impact of Status Culture Participation on the Grades of U.S. High School Students." *American Sociological Review* 47(2):189–201. https://doi.org/10.2307/2094962

DiMaggio, Paul. 1997. "Culture and Cognition." *Annual Review of Sociology* 23(1):263–287. https://doi.org/10.1146/annurev.soc.23.1.263

DiPrete, Thomas A., and Claudia Buchmann. 2013. *The Rise of Women: The Growing Gender Gap in Education and What It Means for American Schools*. New York: Russell Sage Foundation.

DiPrete, Thomas A., and Gregory M. Eirich. 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, Douglas B., and Dennis J. Condron. 2016. "Fifty Years since the Coleman Report: Rethinking the Relationship between Schools and Inequality." *Sociology of Education* 89(3):207–2020. https://doi.org/10.1177/0038040716651676

Dziak, John J., Linda M. Collins, and Aaron T. Wagner. 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* 128(6):1809–1840. https://doi.org/10.1086/724865

Evans, David, and Fei Yuan. 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. https://doi.org/10.1596/1813-9450-8752

Farkas, George. 2003. "Cognitive Skills and Non-cognitive Traits and Behaviors in Stratification Processes." *Annual Review of Sociology* 29:541–562. https://doi.org/10.1146/annurev.soc.29.010202.100023

Fazio, Russell H., Javier A. G. Samayoa, Shelby T. Boggs, and Jesse Ladanyi. 2023. "Implicit Bias: What Is It?" In *The Cambridge Handbook of Implicit Bias and Racism*, edited by Jon A. Krosnick, H. Tobias, and A.L. Scott, Cambridge: Cambridge University Press.

Ferman, Bruno, and Luiz F. Fontes. 2022. "Assessing Knowledge or Classroom Behavior? Evidence of Teachers' Grading Bias." *Journal of Public Economics*. *216*:104773. https://doi.org/10.1016/j.jpubeco.2022.104773

Ferrara, Alessandro. 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

Fiske, Susan T., Monica Lin, and Steven L. Neuberg. 2018. "The continuum model: Ten years later." *Social cognition*, 41-75. https://doi.org/10.4324/9781315187280-3

Foley, William. 2023. "Status Beliefs Negatively Affect Expected University Attainment of Lower Class Students." *Education Inquiry*. https://doi.org/10.1080/20004508.2023.2296143

Foschi, Martha. 2000. "Double Standards for Competence: Theory and Research." *Annual Review of Sociology* 26:21–42. https://doi.org/10.1146/annurev.soc.26.1.21

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

Ganzeboom, Harry B. G., and Donald J. Treiman. 1996. "Internationally Comparable Measures of Occupational Status for the 1988 International Standard Classification of Occupations." *Social Science Research* 25:201–239. https://doi.org/10.1006/ssre.1996.0010

Geven, Sara, Øyving Wiborg, Rachel E. Fish, and Herman G. van de Werfhorst. 2021. "How Teachers Form Future Expectations for Students: A Comparative Factorial Survey Experiment in New York, Amsterdam, and Oslo." *Social Science Research* 100:102599. https://doi.org/10.1016/j.ssresearch.2021.102599

Gil-Hernández, Carlos J., and Pablo Gracia. 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. https://doi.org/10.4054/DemRes.2018.38.23

Gil-Hernández, Carlos J., Leire Salazar, Jonatan Castaño Muñoz, and 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

Glock, Sabine, and Sabine Krolak–Schwerdt. 2014. "Stereotype Activation Versus Application: How Teachers Process and Judge Information About Students from Ethnic Minorities and with Low Socioeconomic Background." *Social Psychology of Education* 17:589–607. https://doi.org/10.1007/s11218-014-9266-6

Glock, Sabine, and Florian Klapproth. 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

Goldthorpe, John H. 2007. "Cultural Capital": Some Critical Observations." *Sociologica* 1(2):1-23.

Gortázar, Lucas, David Martínez de Lafuente, and Ainhoa Vega–Bayo. 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, Anthony G., and Mahzarin R. Banaji. 1995. "Implicit Social Cognition: Attitudes, Self–Esteem, and Stereotypes." *Psychological Review* 102(1):4. https://doi.org/10.1037/0033-295X.102.1.4

Greenwald, Anthony G., and Linda Hamilton Krieger. 2006. "Implicit Bias: Scientific Foundations." *California Law Review* 94(4):945–967. https://doi.org/10.2307/20439056

Hainmueller, Jens, Daniel J. Hopkins, and Teppei Yamamoto. 2014. "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments." *Political Analysis* 22(1):1–30. https://doi.org/10.1093/pan/mpt024

Hanna, Rema N., and Leigh L. Linden. 2012. "Discrimination in Grading." *American Economic Journal: Economic Policy* 4(4):146–168. https://doi.org/10.1257/pol.4.4.146

Heath, Anthony, and Yaël Brinbaum. 2007. "Guest Editorial: Explaining Ethnic Inequalities in Educational Attainment." *Ethnicities* 7(3):291–304. https://doi.org/10.1177/1468796807080230

Hjorth–Trolle, Anders, 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, Anders, Anders Hjorth–Trolle, and Mads Meier Jæger. 2019. "Signals, Educational Decision–Making, and Inequality." *European Sociological Review* 35(4):447–460. https://doi.org/10.1093/esr/jcz010

Homuth, Christoph, Johannes Thielemann, and Sebastian E. Wenz. 2023. "Measuring Elementary School Teachers' Stereotypes in the NEPS SC2." NEPS Survey Paper No. 108. *Leibniz Institute for Educational Trajectories, National Educational Panel Study*.

INE (Instituto Nacional de Estadística). 2020. "Encuesta de Inserción Laboral de los Titulados Universitarios EILU–2019." Madrid: Instituto Nacional de Estadística.

INE (Instituto Nacional de Estadística). 2023a. "Estadística del Padrón Continuo 2011." Madrid: Instituto Nacional de Estadística.

INE (Instituto Nacional de Estadística). 2023b. "Estadística de Nacimientos 2011." Madrid: Instituto Nacional de Estadística.

INEE (Instituto Nacional de Evaluación Educativa). 2010. *Evaluación general de diagnóstico 2009. Educación primaria. Cuarto curso. Informe de resultados*. Madrid: Ministerio de Educación.

Jæger, Mads Meier. 2022. "Cultural Capital and Educational Inequality: An Assessment of the State of the Art." Pp. 121–134 in *Handbook of Sociological Science*, edited by K. Gërxhani, N. D. de Graaf, and W. Raub. Edward Elgar Publishing. https://doi.org/10.4337/9781789909432.00015

Jæger, Mads Meier, and Richard Breen. 2016. "A Dynamic Model of Cultural Reproduction." *American Journal of Sociology* 121(4):1079–1115. https://doi.org/10.1086/684012

Jæger, Mads Meier, and Stine Møllegaard. 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, Mads Meier, Rikke Haudrum Rasmussen, and Anders Holm. 2023. "What Cultural Hierarchy? Cultural Tastes, Status, and Inequality." *The British Journal of Sociology* 74(3):402–418 https://doi.org/10.1111/1468-4446.13012

Jæger, Mads Meier, and Mikkel Haderup Larsen. 2024. "From Metallica to Mozart: Mapping the Cultural Hierarchy of Lifestyle Activities." *Sociological Science* 11: 413-438. https://doi.org/10.15195/v11.a15

Jennings, Jennifer L., and Thomas A. DiPrete. 2010. "Teacher Effects on Social and Behavioral Skills in Early Elementary School." *Sociology of Education* 83(2):135–159. https://doi.org/10.1177/0038040710368011

Kisfalusi, Dorottya, Béla Janky, and Károly Takács. 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):745-780. https://doi.org/10.1177/0272431618791278

Kisfalusi, Dorottya, Béla Janky, and Károly Takács. 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

Krkovic, Katarina, Samuel Greiff, Sirkku Kupiainen, Mari–Pauliina Vainikainen, and Jarkko Hautamäki. 2014. "Teacher Evaluation of Student Ability: What Roles Do Teacher Gender, Student Gender, and Their Interaction Play?" *Educational Research* 56(2):244–257. https://doi.org/10.1080/00131881.2014.898909

Krolak–Schwerdt, Sabine, 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* 86(4):515–529. https://doi.org/10.1080/00220973.2017.1370686

Lamont, Michèle, and Annette Lareau. 1988. "Cultural Capital: Allusions, Gaps and Glissandos in Recent Theoretical Developments." *Sociological Theory* 6:153–68. https://doi.org/10.2307/202113

Lamont, Michèle, and Mario L. Small. 2008. "How Culture Matters: Enriching Our Understandings of Poverty." In *The Colors of Poverty*: *Why Racial and Ethnic Disparities Persist*, edited by D. Harris and A. Lin, 76–102. New York, NY: Russell Sage Foundation.

Lamont, Michèle, Stefan Beljean, and Matthew Clair. 2014. "What Is Missing? Cultural Processes and Causal Pathways to Inequality." *Socio–Economic Review* 12(3):573–608. https://doi.org/10.1093/ser/mwu011

Lareau, Annette. 2011. *Unequal Childhoods: Class, Race, and Family Life. With an Update a Decade Later*. Oakland, CA: University of California Press. https://doi.org/10.1525/9780520949904

Lehmann–Grube, Sabine K., Anita Tobisch, and Markus Dresel. 2023. "Changing Preservice Teacher Students' Stereotypes and Attitudes and Reducing Judgment Biases Concerning Students of Different Family Backgrounds: Effects of a Short Intervention." *Social Psychology of Education*. https://doi.org/10.1007/s11218-023-09862-3

Lievore, Ilaria, 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. https://doi.org/10.1080/01425692.2022.2122942

Lizardo, Omar, and Sara Skiles. 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, Georg, Irena Kogan, Sarah Gentrup, and Cornelia Kristen. 2024. "Non–Native Accents Among School Beginners and Teacher Expectations for Future Student Achievements." *Sociology of Education* 97(1):76-96. https://doi.org/10.1177/00380407231202978

Marcenaro–Gutiérrez, Oscar D., and Anna Vignoles. 2015. "A Comparison of Teacher and Test–Based Assessment for Spanish Primary and Secondary Students." *Educational Research* 57(1):1–21. https://doi.org/10.1080/00131881.2014.983720

Martínez de Lafuente, David. 2021. "Cultural Assimilation and Ethnic Discrimination: An Audit Study with Schools." *Labour Economics* 72:102058. https://doi.org/10.1016/j.labeco.2021.102058

Melamed, David, Christopher W. Munn, Leanne Barry, Bradley Montgomery, and Oneya F. Okuwobi. 2019. "Status Characteristics, Implicit Bias, and the Production of Racial Inequality." *American Sociological Review* 84(6):1013–1036. https://doi.org/10.1177/0003122419879101

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

Miles, Andrew, Raphaël Charron–Chénier, and Cyrus Schleifer. 2019. "Measuring Automatic Cognition: Advancing Dual–Process Research in Sociology." *American Sociological Review* 84(2):308–33. 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, Gregory, and Philip E. Tetlock. 2017. "Popularity as a Poor Proxy for Utility: The Case of Implicit Prejudice." In *Psychological Science Under Scrutiny: Recent Challenges and Proposed Solutions*, edited by S. O. Lilienfeld and I. D. Waldman, pp. 164–195. Wiley Blackwell https://doi.org/10.1002/9781119095910.ch10

Owens, Jayanti. 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

Petzold, Knut. 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, Edmund S. 1972. "The Statistical Theory of Racism and Sexism." *American Economic Review* 62(4):659–661.

Pit–ten Cate, Ineke M., and Sabine Glock. 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

Quinn, David 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. https://doi.org/10.3102/0162373720932188

Ridgeway, Cecilia L. 2014. "Why Status Matters for Inequality." *American Sociological Review* 79(1):1–16. https://doi.org/10.1177/0003122413515997

Salza, Guido. 2022. "Equally Performing, Unfairly Evaluated: The Social Determinants of Grade Repetition in Italian High Schools." *Research in Social Stratification and Mobility* 77:100676. https://doi.org/10.1016/j.rssm.2022.100676

Schuessler, Julian, and Markus Freitag. 2020. "Power Analysis for Conjoint Experiments." *SocArXiv*. December 16. doi:10.31235/osf.io/9yuhp.

Schuessler, Julian, and Kim M. Sønderskov. 2023. "Compensating Discrimination: Behavioral Evidence from Danish School Registers." *SocArXiv*. July 20. doi:10.31235/osf.io/5zm87.

Skopek, Jan, and Giampiero Passaretta. 2021. "Socioeconomic Inequality in Children's Achievement from Infancy to Adolescence: The Case of Germany." *Social Forces* 100(1):86–112. https://doi.org/10.1093/sf/soaa093

Starck, Justin G., Travis Riddle, Stacey Sinclair, and Natasha Warikoo. 2020. "Teachers Are People Too: Examining the Racial Bias of Teachers Compared to Other American Adults." *Educational Researcher* 49(4):273–284. https://doi.org/10.3102/0013189X20912758

Stefanelli, Alberto, and Martin Lukac. 2020. "Subjects, Trials, and Levels: Statistical Power in Conjoint Experiments." *SocArXiv*. November 18. doi:10.31235/osf.io/spkcy.

Südkamp, Anna, Johanna Kaiser, and Jens Möller. 2012. "Accuracy of Teachers' Judgments of Students' Academic Achievement: A Meta–Analysis." *Journal of Educational Psychology* 104(3):743–762. https://doi.org/10.1037/a0027627

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

Tilcsik, András. 2020. "Statistical Discrimination and the Rationalization of Stereotypes." *American Sociological Review* 86(1):93–122. https://doi.org/10.1177/0003122420969399

Timmermans, Anneke C., Hans Kuyper, and Geertje van der Werf. 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. https://doi.org/10.1111/bjep.12087

Timmermans, Anneke C., Hester de Boer, Hilda T. A. Amsing, and Marieke P. C. van der Werf. 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. https://doi.org/10.1002/berj.3470

Van de Werfhorst, Herman G. 2010. "Cultural Capital: Strengths, Weaknesses, and Two Advancements." *British Journal of Sociology of Education* 31(2):157–169. https://doi.org/10.1080/01425690903539065

van Huizen, Thomas, Madelon Jacobs, and Matthijs Oosterveen. 2024. "Teacher Bias or Measurement Error?" arXiv:2401.04200 [econ.EM]. DOI: 10.48550/arXiv.2401.04200.

Wenz, Sebastian E., and Kerstin Hoenig. 2020. "Ethnic and Social Class Discrimination in Education: Experimental Evidence from Germany." *Research in Social Stratification and Mobility* 65:100461. https://doi.org/10.1016/j.rssm.2019.100461

Zanga, Giulietta, and Elena De Gioannis. 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

Zhu, Maria. 2024. "New Findings on Racial Bias in Teachers' Evaluations of Student Achievement." *IZA Discussion Paper* No. 16815. Available at SSRN: https://ssrn.com/abstract=4736400 or http://dx.doi.org/10.2139/ssrn.4736400

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 thank Lilian Weikert, William Foley, Zbigniew Karpiñski, David Martínez de Lafuente, Alberto López, and Mario Spiezio for their valuable feedback and support. We also thank the participants at the following venues where we presented earlier versions of the article: the 'Experiments on Social Inequality' Workshop at Sciences Po-LIEPP, the 'Colloquium of the Migration and Diversity Department' at the WZB Berlin Social Science Centre, the ECSR Thematic Conference 'Effort and Social Inequality' and '2024 IC3JM Conference' at Carlos III-Juan March Institute of Social Sciences, the FES 'Inequality and Social Stratification Committee Workshop' in Oviedo, the 'Education and Social Inequalities Seminar' at University of Sevilla, the 'CLIC Seminar Series' at the European University Institute, the SISEC conference in Cagliari, and the FES National Congress in Sevilla.

Carlos J. Gil-Hernández*: Department of Statistics, Computer Science, Applications, University of Florence. *Corresponding author, E-mail: carlos.gil@unifi.it

Irene Pañeda-Fernández: WZB Berlin Social Science Center. E-mail: irene.paneda@wzb.eu

Leire Salazar: Institute for Public Goods and Policies, Consejo Superior de Investigaciones Científicas. E-mail: leire.salazar@cchs.csic.es

Jonatan Castaño Muñoz: Departamento de Didática y Organización Educativa, Universidad de Sevilla. E-mail: jcastanno@us.es