sociological science

Supplement to:

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: XX-YY.

ONLINE SUPPLEMENT

Part A.

Experimental Set Up: Pre-tests, Pre-registration, and Timeline

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

Table S.1. Experiment Timeline

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

Part B.

Target population: Pre-service teachers

Our population of interest comprises students enrolled in any grade of the 4-year BA Degree in Primary Education in Spain. Holding this degree is a legal requisite to work as a teacher in public elementary schools. According to Spanish administrative data, in 2019/2020, only 9.8% of the students enrolled in the first grade of this BA Degree dropped out, and 3.2% enrolled in another degree (Ministerio de Universidades 2023). As shown in Table 1, our sample attended, on average, the third grade. Furthermore, according to a Spanish survey with a representative sample of the graduates in Primary Education (INE 2020), in 2019, 5 years after graduation, 82% of the Primary Education graduates were employed, of those 76% of those as teachers (ISCO 22-23), 12% unemployed and 6% inactive (70% studying for the teachers' entry exam). Thus, most of our experimental sample of college students will eventually become teachers. We are precisely interested in assessing their potential biases before entering into service.

Data Collection Protocols and Ethics

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

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

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

Table S.2. Population, sample, and response rates

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

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

Part C.

Power Analysis

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

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

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

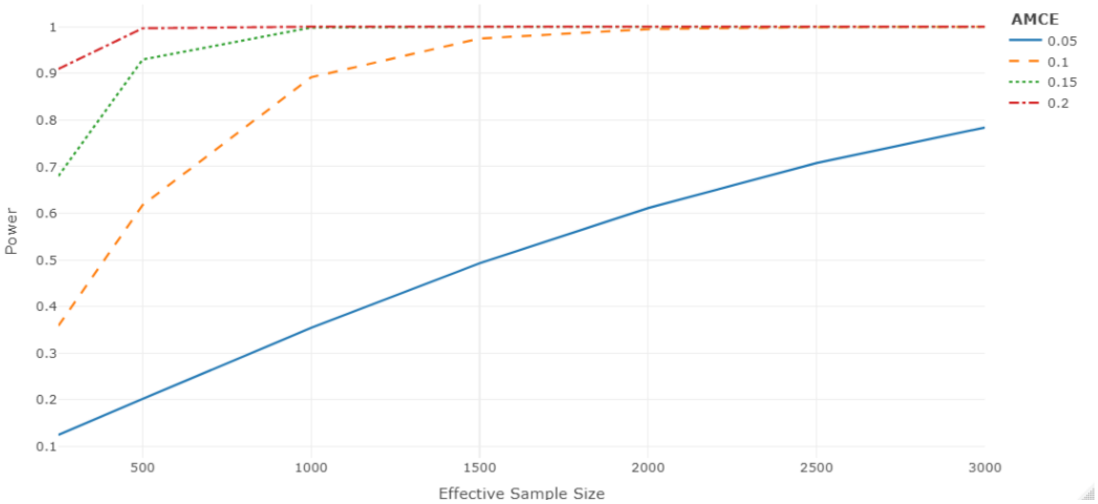
Given that the range of the outcome is different, that they do not find large or statistically significant main SES effects, that their study is most likely underpowered, and that we are not looking at proportions but at mean values (Auspurg and Hinz 2015:33), we find it rather challenging to base calculations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

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

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

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

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



Part D.

Essay Quality Validation and Implementation

Article's Table 3 shows that in-service teachers assigned a 5.5 (SD=1.4) average grade to the bad essay and 8.9 (SD=1.1) to the good essay on a 1-to-10 scale, where 1 is the lowest and 10 is the highest grade, following real grading practice, with a joint mean at 7.2 (SD=2.1). Figure S.2 shows that the distributions of the pre-test (in-service teachers) and experiment (pre-service teachers) essay grades largely overlap. Moreover, we asked teachers to assess the essays' degree of credibility (i.e., written by a 6th grader) and guess the writer's gender. About 60% of respondents reported the essay as credible, and about 70% could not say if a boy or a girl wrote it.

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

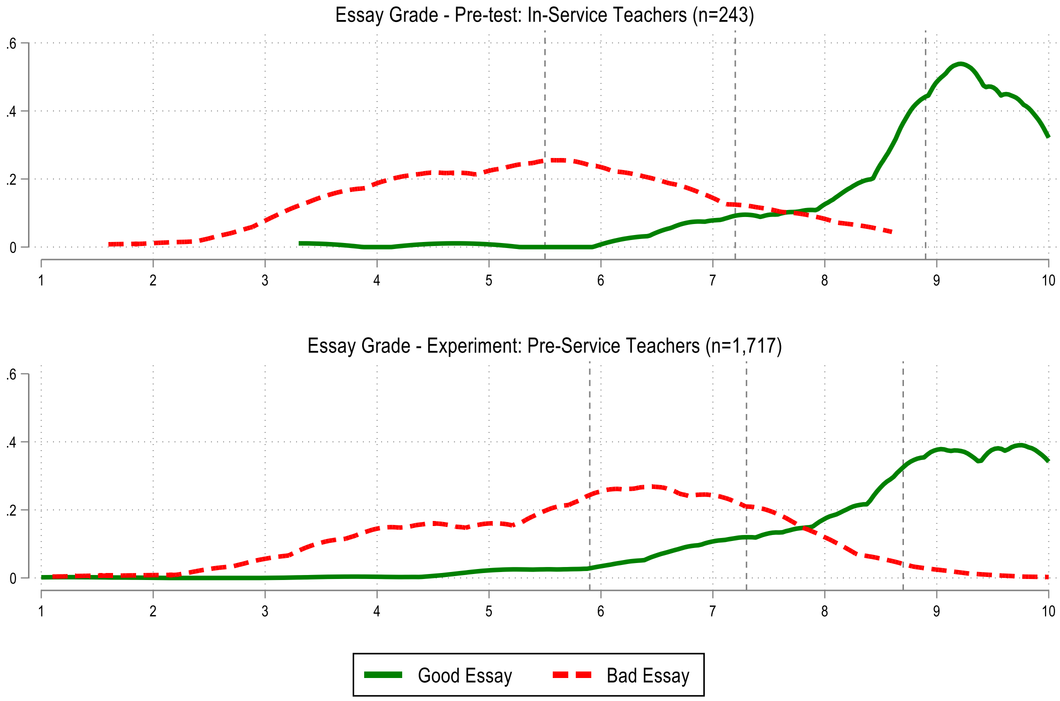


Table S.3. Essay-screen instructions and essay by objective quality, cultural capital, and parental SES signals (in Spanish)

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

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

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

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

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

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

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

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

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

Part E.

Cultural Capital: Signal and Instrument Validation

Table S.4. **Low** / **High** cultural capital signals embedded in the essay (in Spanish)

High Quality Essay

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

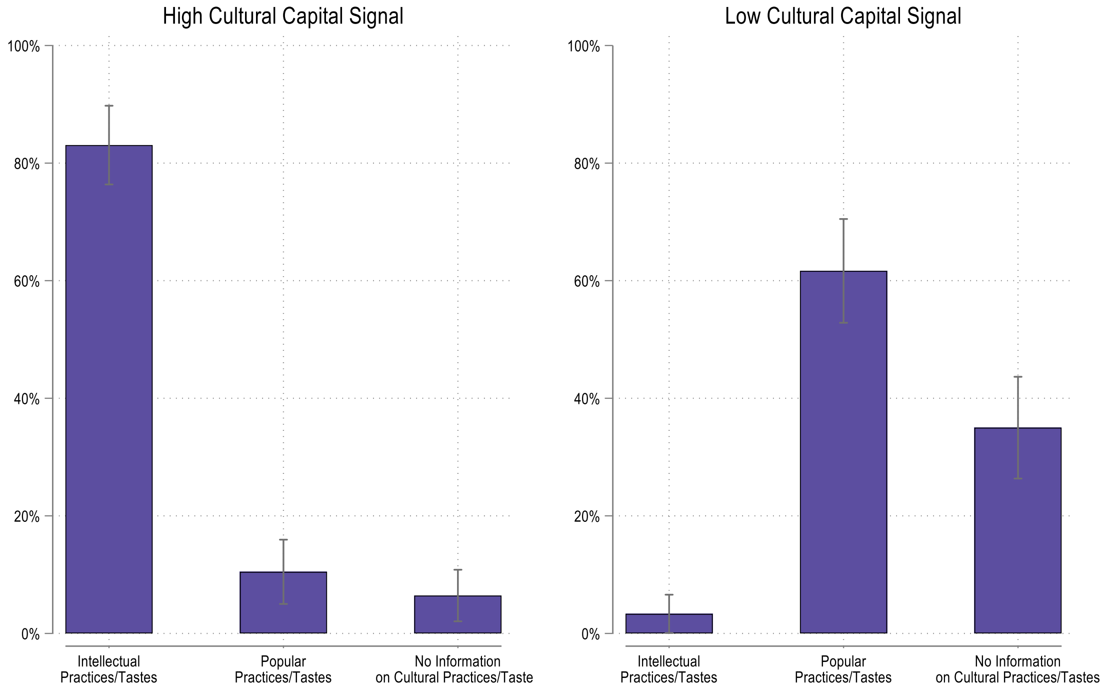
Low Quality Essay

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

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

To ensure that the embodied cultural capital signals shown in Table S.4. are actually perceived as highbrow or lowbrow culture by respondents, in our pre-test with 243 in-service elementary education teachers we asked participants to evaluate which kind of information about the cultural practices and tastes of the student and their family the abovementioned cultural capital indicators suggested to them: (1) intellectual cultural practices and tastes; (2) popular culture practices and tastes; or (3) no information about the student and family cultural practices and tastes. The cultural capital indicators correctly signaled the assumed status hierarchy (Jæger et al. 2023; Childress et al. 2021; Lizardo and Skiles 2009) since, as shown in Figure S.3. below, over 80% of respondents associated the cultural reference to visiting an art museum and knowing an impressionist painter with intellectual, cultural practices and tastes, while 60% associated watching a reality show TV programme with popular cultural practices and tastes. Still, even when about 35% of respondents claimed that the popular culture reference to watching a *trash* TV programme did not convey any information on the cultural practices and tastes of the student and his family, we suspect that a substantial amount of this share might be hiding social desirability bias and avoiding negative labelling since this was asked openly in the pre-test.

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

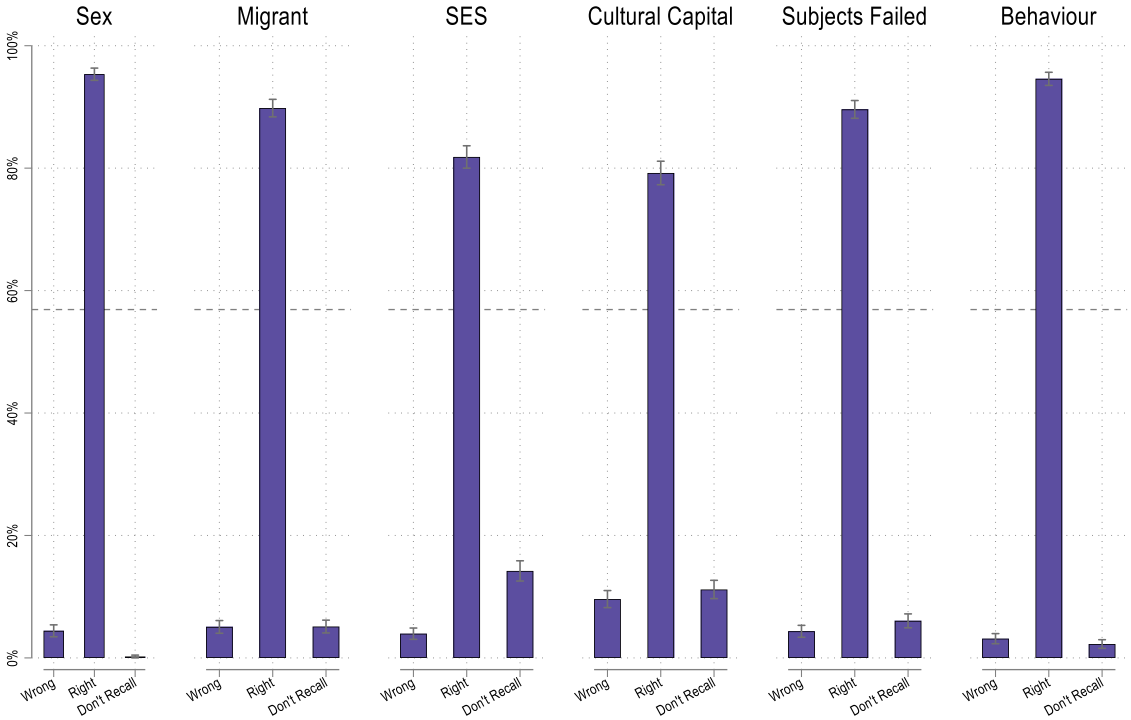


Part F.

Manipulation Checks

We included a post-experimental survey module including several questions as manipulation checks to assess the effectiveness of the study's factorial manipulations or randomized treatments. These checks ensure that the signals, such as cultural capital markers (see above), the student's parental SES, ethnic origin, and gender, along with the students' ability-related factors, are working as intended by being correctly recognized and remembered by the respondents. That is key in our design for causally identifying potential biases in respondents' assessments by the randomized treatments while properly controlling for all the relevant confounders. However, not remembering the factors could also be a proxy for not paying enough attention to that information precisely because the participant might not consider it relevant for the required assessment. Figure S.4 shows that the correct recall of single treatments or factor levels is over 80%, varying from 79% for cultural capital to 95% for gender and behaviour; 57% of respondents correctly recalled all factorial manipulations included. We run robustness checks of all the main analyses on a subsample of respondents correctly recalling all treatments (See Online Supplement Part I, Table S.6.).

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



Part G.

Vignettes Randomization and Distribution

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

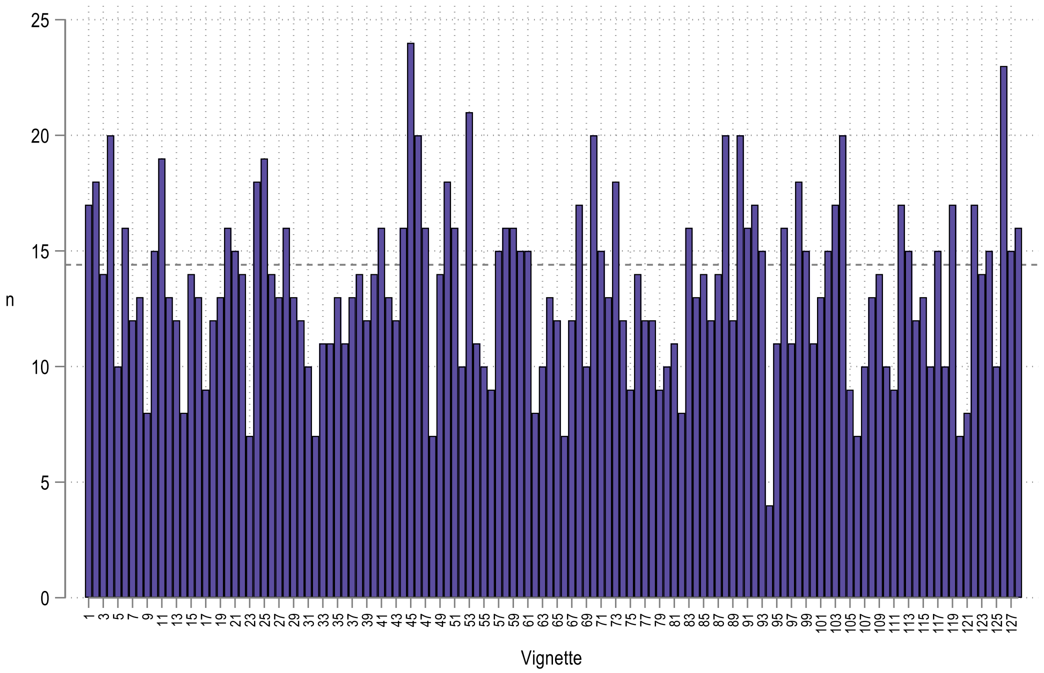
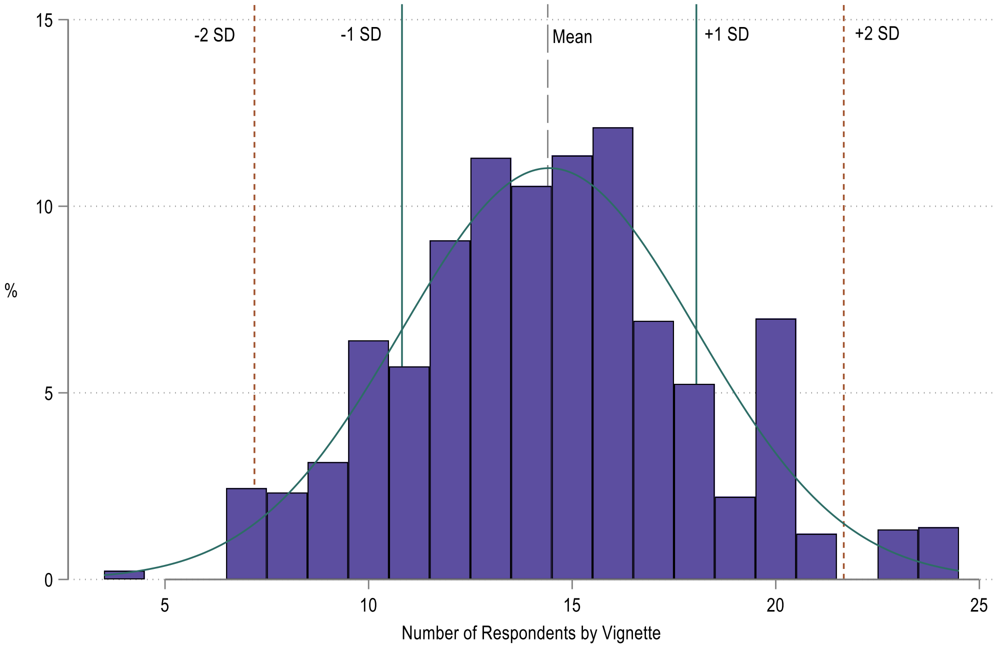


Figure S.6. Distribution of Number of Respondents (n=1,717) by Vignette (n=128)



Part H.

Main Models' Full Output

Table S.5A. OLS main models (M2): Experimental factors AMCEs on educational outcomes

	Outcomes		
Randomized Factors	Essay Grade (1-10)	Grade Retention Recommendation (0-10)	Academic Track Expectations (0-10)
Ascriptive Factors			
Female (Male)	0.12^{+}	-0.13	0.24^{**}
	(0.06)	(0.12)	(0.07)
Native Origin (Moroccan Origin)	-0.20^{**}	0.13	0.19^{*}
	(0.06)	(0.11)	(0.07)
High-SES (Low-SES)	0.03	-0.03	0.20^{*}
	(0.06)	(0.11)	(0.08)
High Cultural Capital (Low CC)	0.20^{***}	-0.09	0.09
	(0.05)	(0.12)	(0.07)
Ability Factors			
Good Essay (Bad Essay)	2.83^{***}	-2.17^{***}	1.31^{***}
	(0.11)	(0.14)	(0.10)
All Subjects Passed (3 Core Subjects Failed)	0.28^{**}	-1.73^{***}	0.46^{**}
	(0.07)	(0.09)	(0.12)
Good Behavior + Effort (Bad Behavior + Effort)	0.27^{**}	-1.03^{***}	1.21^{***}
	(0.08)	(0.10)	(0.10)
Individual Controls	✓	✓	✓
Institution Fixed Effects	✓	✓	✓
Ratio Ability / Ascriptive[a]	8.15	17.78	5.56
Observations	1,717	1,717	1,717
Adjusted R^2	0.52	0.25	0.19
Root Mean Square Error	1.37	2.55	1.95

Notes: Clustered standard errors by institutions in parentheses, individual-level controls: year of birth, gender, country of birth, parental country of birth, parental highest education, BA Degree enrollment grade, grade retention in primary and/or secondary school.

[a] *Ratio Ability / Ascriptive* is calculated by dividing the average absolute effect size of the three ability factors by the average absolute effect size of the four ascriptive factors. Two-tailed t-tests: $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table S.5B. Main OLS models (M1 and M2)

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

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

Part I.

Robustness Checks

We run several pre-registered robustness checks that support the main findings. Firstly, in Table S.6., we replicate analyses in a subsample of those respondents who correctly recalled all treatment levels in the manipulation checks (56.9%; n=977). Second, as a deviation from the pre-analysis plan, we generated calibration weights using raking estimators to adjust for the population shares of the main individual-level socio-demographic variables (see Table S.7. below). Third, given that our primary outcomes are significantly non-normally distributed (see Table 4), we dichotomize the outcomes below/above the median combined with linear probability models (LPM) in Table S.8. Fourth, for the outcome on grade retention, in the appendix Figure S.9., we display a heterogenous model (M2) by the number of failed subjects (none or three core subjects) to mitigate the skewness in the joint distribution and test for a more realistic setting.

Table S.6. Manipulation check: main model M2 and M2 among the subsample correctly recalling all signals (M2 | Signals)

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

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

Table S.7. Main models without and with weighting by population socio-demographics

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

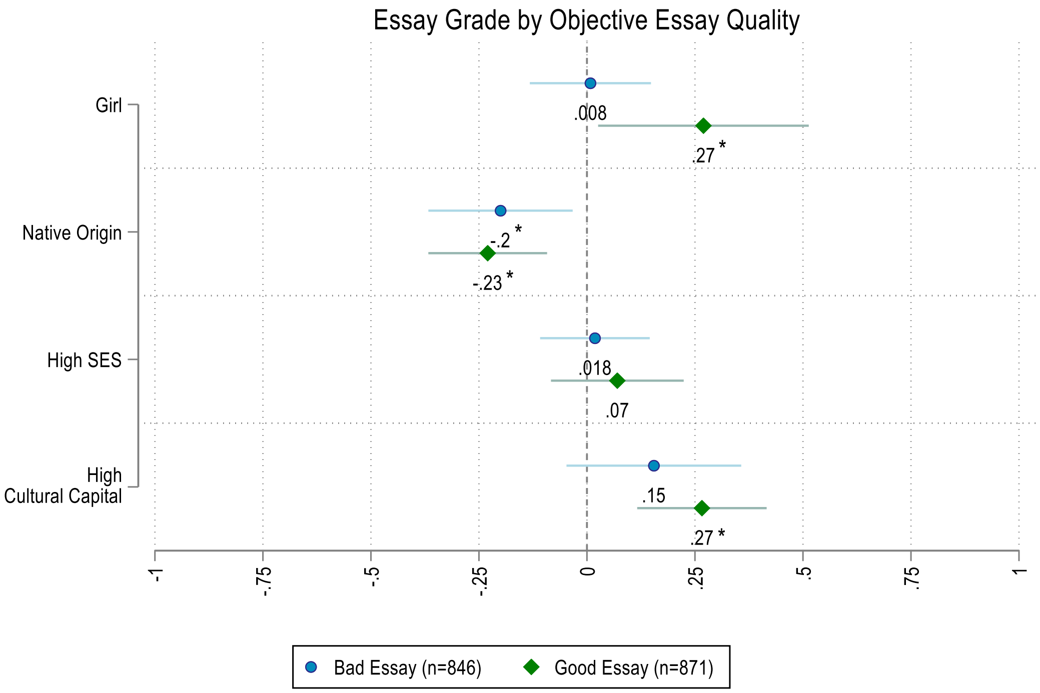
Notes: Clustered standard errors by institutions in parentheses; $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table S.8. Main OLS models and LPM with dummy outcomes (below/above median)

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

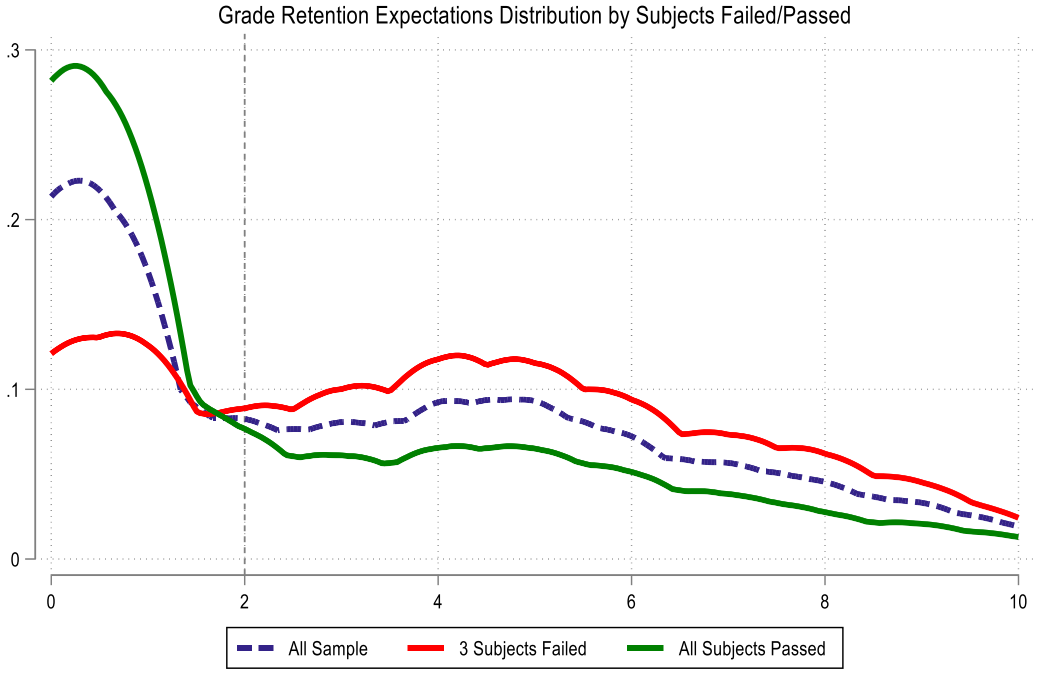
Notes: Clustered standard errors by institutions in parentheses, individual-level controls: year of birth, gender, country of birth, parental country of birth, parental highest education, BA Degree enrollment grade, grade retention in primary and/or secondary school. Two-tailed t-tests: $^{+}$ $p < 0.10$,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

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



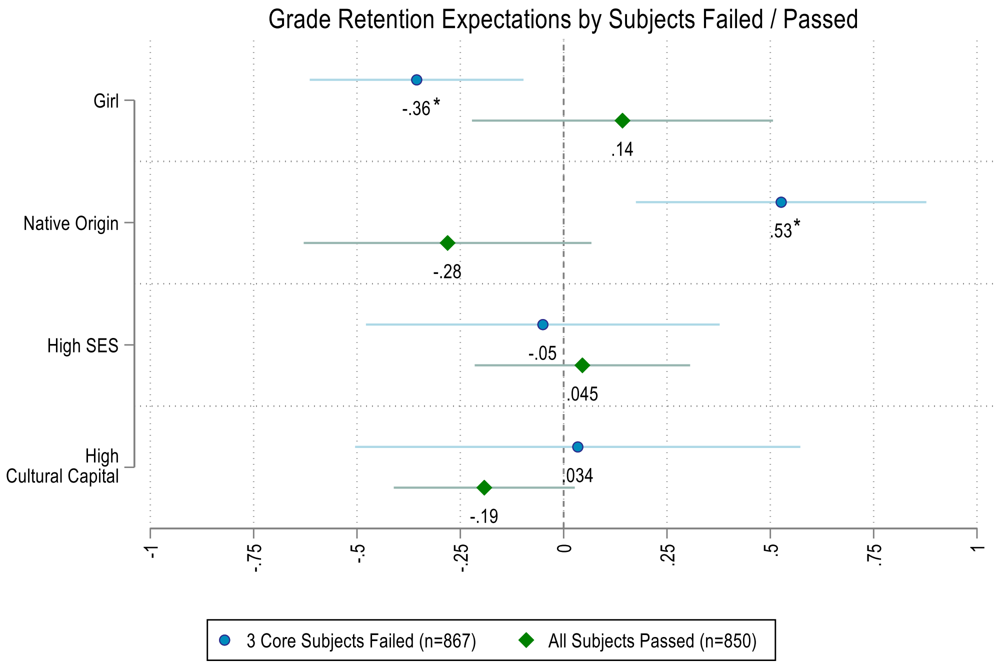
Notes: *p-value < 0.05; Controls: institution-FE; respondents' characteristics; ability factors.

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



Notes: Median all sample = 2

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



Notes: *p-value < 0.05; Controls: institution-FE; respondents' characteristics; ability factors.

Part J.

Mechanisms

In a third set of models (M3), we test whether the teachers' perception of parental support available for student's education (0-10 scale) is a mechanism of ascribed factors. Parental support was asked with the following question: *Considering the information in the student's file, how much interest and support do you think the family shows in the student's education? On the scale, 0 means no interest or support, and 10 means a lot of interest and support*. We (1) run OLS models of the factors on parental support (Table S.9.), and (2) use the Karlson-Holm-Breen (KHB) decomposition (Table S.10.). Parental support only mediates or confound the effect of cultural capital on essay grading at 36% (p-value<0.1). This finding backs the interpretation of statistical discrimination in long-term expectations instead of being an educated guess that assumes low-SES or migrant-origin students will count on less parental support during their prospective education.

Table S.9. Mechanisms: Ascriptive Factors on Parental Support

	Parental Support (0-10)
	M3
Female	0.0526
	(0.109)
Spanish Origin	-0.0852
	(0.116)
High-SES	0.142
	(0.122)
High Cultural Capital	0.497***
	(0.087)
Good Essay	1.115***
	(0.088)
All Subjects Passed	0.621***
	(0.104)
Good Behavior + Effort	2.180***
	(0.094)
Institution FE	✓
Individual Controls	✓
Observations	1,717
Adjusted R²	0.240

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

Table S.10. Mechanisms: KHB Linear Models on Confounding / Mediation of Parental Support on Outcomes by Ascriptive Factors

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

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

Part K.

Hypothesis Testing

Table S.11. Coefficient differences across outcome models

	$\Delta\beta_{y3-y1}$ (Z-Acad. Track Expectations – Z-Essay Grade)		$\Delta\beta_{y3-y2}$ (Z-Acad. Track Expectations – Z-No Grade Retention)	
Randomized Factors	Δβ	z	Δβ	z
Ascriptive Factors				
Girl	0.05	1.38	0.07	1.44
	(0.04)		(0.05)	
Native Origin	0.19***	6.4	0.13***	3.59
	(0.03)		(0.04)	
High-SES	0.08	1.25	0.08^{+}	1.83
	(0.06)		(0.05)	
High Cultural Capital	-0.06^{+}	-1.65	0.01	0.29
	(0.04)		(0.04)	
Ability Factors				
Good Essay	-0.83***	-17.07	-0.13**	-3.28
	(0.05)		(0.04)	
All Subjects Passed	0.07^{+}	1.74	-0.37***	-6.98
	(0.04)		(0.05)	
Good Behavior + Effort	0.42***	9.2	0.21***	4.95
	(0.05)		(0.04)	

Notes: Coefficients difference test from different models with a two-tailed z-test using seemingly unrelated regressions (Clogg, Petkova, and Haritou 1995). Reversed scale for grade retention recommendation outcome to accurately estimate coefficient differences. Outcomes in z-scores for scale comparability. Clustered standard errors by institutions in parentheses. Two-tailed z-tests: $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Part L

Benchmarking

On average, as shown in Table S.12., we reported average effect sizes (Cohen's D ≈ 0.1 or 10% an SD) close to previous observational studies in Denmark (Schuessler and Sønderskov 2023) and Italy (Alesina et al. 2018) but considerably smaller than the most comparable observational study in Spain (Gortázar et al. 2022). Thus, observational studies might overestimate teacher bias when not accounting for measurement error in test scores and/or not controlling for non-cognitive ability measures (van Huizen et al. 2024). Ideally, to accurately identify teacher biases with observational data, one should exploit residual differences between fully comparable high-stakes blind test scores and teacher-assigned grades covering the same curricula while controlling for students' socio-emotional skills (Schuessler and Sønderskov 2023; Ferman and Fontes 2022; Bygren 2020).

Are our experimental estimates of teacher bias substantial as an educational inequality mechanism relative to students' ability or inequalities in its formation? As illustrated in Appendix Table S.12., for benchmarking, we calculated ascribed status gaps in a standardized test of Spanish competencies, grade retention, and educational expectations in a nationwide evaluation of 4th graders. For instance, our experimental estimates of teacher bias account for more than 50% of the observed gender gaps across all three outcomes. We can also benchmark our average discrimination effect sizes (0.1 SD) with mean learning gains over a school year (0.15-0.21 SD of literacy ability) or large-scale educational interventions (0.17-0.47 SD) (Evans and Yuan 2019).

Table S.12. Findings summary and benchmarking with observational research

	(1) Observational Teacher Bias[b] (SD)[a]	(2) Experimental Teacher Bias[c] (SD)	(3) Experimental / Observational (2/1)*100	(4) Total Observed Gap[d] (SD)	(5) Experimental / Total Gap (2/4)*100	(6) Hypotheses Validation (2)
			Grading			
Girl	0.27**	0.06^{+} [0.14*]	23 %	0.12***	53 %	✓ (H1a)
High-SES	0.22**	0.02	8 %	0.54***	3 %	x (H1a)
Native Origin	0.14**	-0.10**	-69 %	0.79***	-13 %	**x** (H1a)
High Cultural Capital		0.10***				✓ (H3a)
			Grade Retention			
Girl		-0.04 [-0.36*]		-0.08***	55 %	✓ (H1b)
High-SES		-0.01		-0.24***	4 %	x (H1b)
Native Origin		0.04 [0.53**]		-0.32***	-14 %	**x** (H1b)
High Cultural Capital		-0.03				x (H3b)
			Educational Expectations			
Girl		0.11**		0.14***	78 %	✓ (H1c; H2)
High-SES		0.09*		0.46***	20 %	✓ (H1c; H2)
Native Origin		0.09*		0.01	664 %	✓ (H1c; H2)
High Cultural Capital		0.04				x (H3c)

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

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

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

[d.] Own elaboration with data (n=22,500) from a national evaluation among 4th graders (INEE 2010); High-SES = skilled workers vs. professionals (fathers); native = students with both parents born in Spain vs. 2nd generation Moroccan-origin students with both parents born in Morocco. OLS and LPM on Spanish standardized blind test scores, grade retention in 2nd or 4th grade, and parental expectations (`*What educational level are you hoping for that your child is studying?*´) for their children's educational attainment (1 = university or academic upper-secondary track; 0 = compulsory education or vocational training) with controls for gender, father's occupation, migrant-origin, and month of birth, with clustered standard errors by schools.

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